

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ  
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ  
ВЫСШЕГО ОБРАЗОВАНИЯ  
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЯДЕРНЫЙ УНИВЕРСИТЕТ «МИФИ»  
(НИЯУ МИФИ)

ИНСТИТУТ ЯДЕРНОЙ ФИЗИКИ И ТЕХНОЛОГИЙ  
КАФЕДРА №40 «ФИЗИКА ЭЛЕМЕНТАРНЫХ ЧАСТИЦ»

УДК 53.05, 53.07

**ПОЯСНИТЕЛЬНАЯ ЗАПИСКА  
К БАКАЛАВРСКОЙ ДИПЛОМНОЙ РАБОТЕ  
КЛАССИФИКАЦИЯ ЛЕПТОННЫХ РАСПАДОВ W БОЗОНА  
МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ В P-P  
СТОЛКНОВЕНИЯХ ПРИ  $\sqrt{S} = 13$  ТЕВ В ЭКСПЕРИМЕНТЕ  
ATLAS**

Студент

\_\_\_\_\_ Г. А. Толкачёв

Научный руководитель,  
к.ф.-м.н., доц.

\_\_\_\_\_ Е. Ю. Солдатов

Научный консультант,  
инж.

\_\_\_\_\_ Д. Е. Пономаренко

Москва 2021

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА

**КЛАССИФИКАЦИЯ ЛЕПТОННЫХ РАСПАДОВ W БОЗОНА  
МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ В P-P  
СТОЛКНОВЕНИЯХ ПРИ  $\sqrt{S} = 13$  ТЕВ В ЭКСПЕРИМЕНТЕ  
ATLAS**

Студент	_____ Г. А. Толкачёв
Научный руководитель, к.ф.-м.н., доц.	_____ Е. Ю. Солдатов
Научный консультант, инж.	_____ Д. Е. Пономаренко
Рецензент, к.ф.-м.н., с.н.с	_____ Ю. Г. Нарышкин
Секретарь ГЭК, к.ф.-м.н.	_____ А. А. Кириллов
Зав. каф. №40, д.ф.-м.н., проф.	_____ М. Д. Скорохватов

# Оглавление

<b>Введение</b>	<b>4</b>
<b>1 Основные сведения</b>	<b>6</b>
1.1 Стандартная модель . . . . .	6
1.2 Лептонная универсальность . . . . .	8
1.3 Современный статус экспериментальной проверки лептонной универсальности . . . . .	9
1.3.1 Измерения в распадах $B$ мезонов . . . . .	9
1.3.2 Измерения в распадах $W$ бозона . . . . .	11
1.4 Машинное обучение . . . . .	15
1.4.1 Классификатор – метод ближайших соседей (kNN) . . . . .	15
1.4.2 Классификатор – лес деревьев решений (BDT) . . . . .	16
1.4.3 Методы оценки классификатора . . . . .	18
1.5 Метод максимального правдоподобия . . . . .	19
1.5.1 Функция правдоподобия . . . . .	20
1.6 Стратегия анализа . . . . .	22
<b>2 Экспериментальная установка</b>	<b>25</b>
2.1 Большой адронный коллайдер . . . . .	25
2.2 Детектор ATLAS . . . . .	27
2.2.1 Система координат . . . . .	29
<b>3 Исследование методов машинного обучения</b>	<b>30</b>
3.1 Используемые данные . . . . .	30
3.1.1 Отбор событий . . . . .	30
3.1.2 Сравнение кинематических распределений мюонного и электронного канала . . . . .	31

3.2	Подбор оптимального классификатора . . . . .	34
<b>4</b>	<b>Исследование отклика классификатора BDT</b>	<b>38</b>
4.1	Использованные данные . . . . .	38
4.2	Контрольные регионы . . . . .	40
4.2.1	Z регион . . . . .	40
4.2.2	Сигнальный регион . . . . .	41
4.2.3	Оценка КХД фона . . . . .	44
4.2.4	Псевдо-W регион . . . . .	45
4.3	Обучение классификатора BDT . . . . .	46
4.4	Результат классификации событий . . . . .	48
<b>5</b>	<b>Измерение лептонной универсальности</b>	<b>51</b>
5.1	Оптимизация бинирования BDT . . . . .	51
5.2	Учет систематических погрешностей . . . . .	53
5.3	Результаты . . . . .	56
	<b>Заключение</b>	<b>61</b>
	Список использованных источников . . . . .	63

# Введение

Исследования в физике элементарных частиц привели к созданию теории взаимодействия частиц на субъядерном уровне, которую принято называть Стандартной моделью. В Стандартной модели существует три поколения лептонов. Согласно лептонной универсальности предполагается, что три поколения лептонов во всех процессах должны вести себя одинаково. В рамках данного предположения отношение отношений сечений лептонного распада  $W$  бозона для любой пары лептонных каналов распада должно быть равно единице.

В анализе данных с LEP имеются различия между теоретическим предсказанием Стандартной модели и экспериментальными измерениями[1]. А именно, существует указание на возможное отклонение в отношении отношений сечения  $\mathcal{B}(W \rightarrow \tau\nu \rightarrow \mu\nu\nu)/\mathcal{B}(W \rightarrow \mu\nu)$  двух процессов лептонного распада  $W$  бозона (см. раздел 1.4). Однако, данное измерение не обладает достаточной статистической силой для того, чтобы дать однозначный ответ. Из-за чего данное измерение необходимо провести с большей статистической силой.

На эксперименте ATLAS во время второго сеанса набора данных рождалось колоссальное число  $W$  бозонов. Даже не смотря на достаточно высокий триггерный порог детектора, эксперимент ATLAS накапливает на порядки больше статистики, чем было набрано на LEP. Поэтому можно ожидать большую статистику распадов  $W$  бозонов. Однако, имеется проблема, заключающаяся в схожести сигнатуры событий распадов  $W$  бозона в мюон или электрон и в  $\tau$  лептон (см. рисунок 1.6.1). Это происходит из-за того, что сам  $\tau$  лептон распадается на лептоны первого или второго поколения, поэтому эти процессы очень близки кинематически к друг другу. Для проведения измерения отношения отношений сечений необходимо подобрать такую переменную, которая будет максимально отражать

разницу в кинематике этих распадов. Разделение можно произвести, например, с использованием кинематической переменной поперечной массы  $m_T$ . Или же можно использовать переменную, которая комбинирует в себе несколько кинематических переменных.

Анализ, в рамках которого проводится работа, является актуальным для поиска отклонения от предсказаний Стандартной модели. Если в результате данного анализа полученный результат совпадет с теоретическим предсказанием, то это может свидетельствовать о подтверждении лептонной универсальности. В противном случае, если данный анализ подтвердит возможное отклонение в анализе данных с LEP, то это укажет на однозначное существование Новой физики. Возможное отклонение от предсказаний Стандартной модели можно объяснить несколькими способами. Например, одно из возможных объяснений требует введения новой частицы, по-разному взаимодействующей с различными лептонами. Возможно, это новый класс частиц, например, лептокварки[2] или же тяжелый аналог  $Z$  бозона[3].

*Целью данной дипломной работы является выбор оптимальной модели машинного обучения, с помощью которой можно провести классификацию данных. В результате классификации данных можно получить переменную, которая объединит кинематические переменные, используемые при классификации. Предполагается, что на основе данной переменной удастся лучше разделить кинематически очень схожие события. Таким образом, в дальнейшем, можно будет провести более точное измерение отношения отношений сечений лептонных распадов  $W$  бозона.*

# Глава 1

## ОСНОВНЫЕ СВЕДЕНИЯ

В данной главе дана вводная теоретическая информация о Стандартной модели и о лептонной универсальности. Также в этой главе описаны методы машинного обучения, с помощью которых предполагается улучшить результат измерения лептонной универсальности. В главе описан метод максимального правдоподобия, с помощью которого в этой работе производится измерение отношения отношений сечений.

В разделе 1.1 описана Стандартная модель. Далее в разделе 1.3 рассказано о лептонной универсальности, а в разделе 1.4 описаны уже существующие эксперименты по проверке лептонной универсальности. В разделе 1.5 описаны методы машинного обучения, используемые в работе. В разделе 1.5 описан метод максимального правдоподобия и процесс задания функции правдоподобия, а в разделе 1.6 описан процесс измерения отношения отношения сечений с использованием новой переменной отклика классификатора.

### 1.1 Стандартная модель

Стандартная модель (СМ)[4] является одной из важнейших теоретических конструкций в физике элементарных частиц, которая позволяет теоретически предсказать свойства различных процессов в физике элементарных частиц. В рамках Стандартной модели имеется 2 типа элементарных частиц: бозоны и фермионы.

Фермионы имеют полуцелый спин, поэтому описываются статистикой Ферми-Дирака. Сами фермионы делятся на две подгруппы: кварки

и лептоны. Считается, что лептоны это частицы, не участвующие в сильных взаимодействиях и не имеющие внутренней структуры. Все они имеют спин, равный  $1/2$ . Лептоны делятся на электрически заряженные частицы ( $e, \mu, \tau$ ) и электрически нейтральные частицы - нейтрино ( $\nu_e, \nu_\mu, \nu_\tau$ ). Каждому заряженному лептону ставится в соответствие свое нейтрино, вместе с которым они образуют семейство лептонов. Известно, что в Стандартной модели существует три лептонных семейства:

$$(\nu_e, e^-)(\nu_\mu, \mu^-)(\nu_\tau, \tau^-)$$

Все лептонные семейства сохраняются в процессах, описываемых Стандартной моделью.

Кварки являются массивными частицами, имеющие электрический и цветовой заряд, в настоящее время известно шесть различных ароматов кварков:  $u, d, c, s, t, b$ . Кварки, как и лептоны, образуют три семейства:

$$(u, d)(c, s)(t, b)$$

Кварки, как и лептоны, имеют спин  $1/2$ . Все кварки имеют электрический заряд. Электрический заряд  $u, c, t$  кварков, равен  $+2/3$ , а у кварков  $d, s, b$ , равен  $-1/3$  в единицах элементарного заряда. Помимо электрического заряда у кварков имеется цветовой заряд. Цветовой заряд условно обозначают за красный, синий и зеленый. Сильные взаимодействия между кварками обусловлены наличием у кварков этих цветовых зарядов. Помимо сильного взаимодействия кварки могут участвовать в слабых и электромагнитных взаимодействиях. Кварки не могут существовать в несвязанном состоянии, поэтому существуют бесцветные системы, состоящие из нескольких кварков - адроны. Всем частицам Стандартной модели ставится в соответствии античастица, у которых противоположный электрический заряд, но масса и спин такая же.

Помимо фермионов в Стандартную модель входят калибровочные бозоны, являющиеся переносчиками трех фундаментальных взаимодействий, входящих в Стандартную модель. Переносчиком электромагнитного взаимодействия является безмассовый фотон  $\gamma$ . За слабое взаимодействия отвечают  $W^\pm$  и  $Z^0$  бозоны, которые имеют массу. А за сильное взаимодействие

между частицами отвечают 9 безмассовых глюонов. Калибровочные бозоны имеют спин 1, поэтому описываются статистикой Бозе-Эйнштейна.

Для завершения построения Стандартной модели, необходимо ввести бозон Хиггса. Именно за счет взаимодействия с полем Хиггса частицы обретают массу. Стандартная модель построена на локальной калибровочной симметрии

$$SU(3)_c \oplus SU(2)_L \oplus U(1)_Y,$$

спонтанно нарушаемой за счет скалярного Хиггсова поля с ненулевым вакуумным средним до

$$SU(3)_c \oplus U(1)_{em}.$$

После спонтанного нарушения симметрии все фермионы приобретают массу, кроме нейтрино. Приобретают массу бозоны группы  $SU(2)_L \oplus U(1)_Y$ , кроме фотона. Глюоны остаются безмассовыми, т.к. цветовая симметрия не нарушена.

## 1.2 Лептонная универсальность

В рамках Стандартной модели существует три поколения лептонов. Три поколения упорядочены по массе заряженного лептона в диапазоне от 0.511 МэВ для  $e^\pm$  до 105 МэВ для  $\mu^\pm$  и 1777 МэВ для  $\tau^\pm$  [5]. Различия в массе приводят к совершенно разным временам жизни. От стабильного электрона до 2.2 мкс для мюона, и 0.29 пс для тау лептона. Лептоны участвуют в электромагнитных и слабых, но не сильных взаимодействиях, тогда как нейтрино участвуют только в слабом взаимодействии. Стандартная модель предполагает, что эти взаимодействия заряженных и нейтральных лептонов универсальны, т.е. одинаковы для трех поколений. Количество поколений лептонов пока не объяснено в рамках Стандартной модели. Почти все наблюдаемые во вселенной процессы выглядели бы точно так же, если бы существовало только одно поколение лептонов.

## 1.3 Современный статус экспериментальной проверки лептонной универсальности

Для проверки лептонной универсальности выбирают два распада частицы, в одном из которых участвуют лептон одного поколения, а в другом - полностью идентичном лептон другого поколения. Для каждого из них определяется вероятность распада. Согласно Стандартной модели три поколения заряженных и нейтральных лептонов универсальны. При проверке лептонной универсальности предполагают, что разница между тремя поколениями лептонов может быть обусловлена только отличающейся массой этих частиц, в то время как константы слабого и электромагнитного взаимодействия для них одинаковы. Поэтому отношение этих вероятностей двух распадов должно равняться единице.

Измерения лептонной универсальности выполнялись в течении десятилетия на многих экспериментах. Тем не менее до сегодняшнего дня однозначного отклонения от модели не наблюдается с достаточной достоверностью.

### 1.3.1 Измерения в распадах $B$ мезонов

Для поиска отклонений от лептонной универсальности подходят распады, в которых  $b$  - кварк переходит  $s$  - кварк. Примером такого процесса является распад  $B^+$ -мезона на  $K^+$ -мезон и два заряженных лептона  $l^+l^-$ . В рамках Стандартной Модели не существует частицы, которая способна напрямую превратить  $b$  - кварк внутри  $B$ -мезона в  $s$  -кварк внутри  $K$  - мезона. Этот процесс идет внутри мезона с помощью тяжелых виртуальных частиц. Поэтому вероятность данного процесса подавлена. Но раз вклад Стандартной модели так мал, то слабые эффекты гипотетической Новой физики, которые не удастся заметить в других распадах, в данном процессе могут стать существенными.

На рисунке 1.3.1 приведены кварковые диаграммы для распадов  $B^+ \rightarrow K^+l^+l^-$ , отражающие как механизмы СМ, так и влияние гипотетических лептокварков. Если константа для вершины кварк-лептон-лептокварк различна для электронов и мюонов, то соответствующие диаграммы дадут

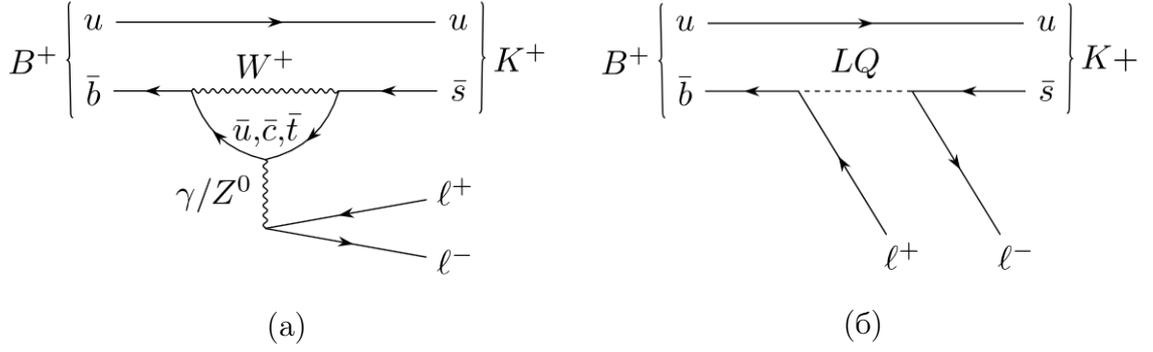


Рисунок 1.3.1 – Один из вкладов Стандартной модели (а) и возможный вклад (б) Новой физики в  $b \rightarrow s$  переходы через промежуточный лептокварк LQ.

различный вклад в вероятность протекания процессов распада с электрон-позитронной и мюон-антимюонной парами в конечном состоянии. Из за чего отношение отношений сечений

$$\mathcal{R}_{K_+} = \mathcal{B}(B^+ \rightarrow K^+ \mu^+ \mu^-) / \mathcal{B}(B^+ \rightarrow K^+ e^+ e^-) \quad (1.1)$$

будет отличаться от единицы.

Одним из экспериментов, который занимается проверкой лептонной универсальности является эксперимент LHCb на Большом адронном коллайдере (см. раздел 2.1). В 2021 году коллаборация LHCb опубликовала статью, посвященную измерению отношения  $\mathcal{R}_{K_+}$  [6]. Результат был получен в результате анализа данных, набранных в ходе полной второй сессии работы Большого адронного коллайдера. При анализе был исследован сигнал в малых областях  $q^2$  квадрата масс лептонной пары. Значение  $R_{K_+}$  измерено в диапазоне  $1.1 < q^2 < 6.0^2$  и равно:

$$\mathcal{R}_{K_+} = 0.846_{-0.041}^{+0.044} \quad (1.2)$$

На рисунке 1.3.2 показано экспериментальные результаты по измерению  $\mathcal{R}_{K_+}$  с экспериментов BaBar [7], Belle [8], а также текущий и предыдущий [9] результата с LHCb. Предыдущий результат измерения был получен с использованием данных, набранных во время второй сессии набора данных в 2015 и 2016 годах и первой сессии набора данных Большого адронного коллайдера. При новом измерении использовалась вся статистика второй сессии набора данных, что примерно вдвое превышает статистику, использованную в предыдущем измерении на LHCb. Сравнивая с результатом

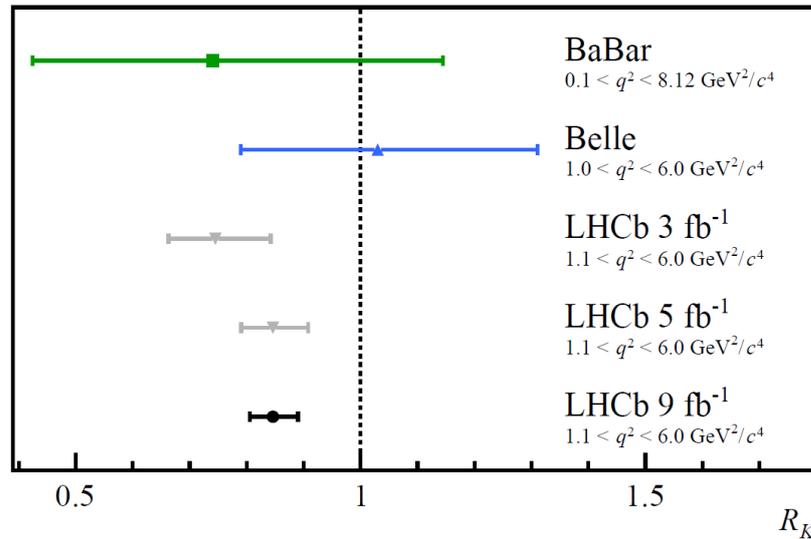


Рисунок 1.3.2 – Результаты проверки лептонной универсальности через отношение рению  $\mathcal{R}_{K^+}$  по данным различных экспериментов: в детекторах BaBar и Belle на электрон-позитронных коллайдерах и в детекторе LHCb на все возрастающей статистике. Черная точка с погрешностями – значение 2021 года. Вертикальный пунктир – предсказание Стандартной модели.

предыдущего измерения  $\mathcal{R}_{K^+}$  можно заметить, что погрешность измерения заметно уменьшилась, но центральное значение осталось прежним. Новый результат отличается от единицы уже на 3.1 стандартных отклонения. Что может говорить о существовании экспериментального указания на нарушение лептонной универсальности в распадах  $B^+$ -мезонов

Новый результат отличается от единицы уже на 3.1 стандартных отклонения. Что может говорить о существовании экспериментального указания на нарушения лептонной универсальности в распадах  $B^+$ -мезонов.

### 1.3.2 Измерения в распадах $W$ бозона

Проверка лептонной универсальности проводится также в распадах  $W$  бозона. Так, измерение лептонной универсальности проводилось на экспериментах ускорителя LEP[10] (ALEPH, DELPHI, L3 и OPAL)[11]. Начиная с 1996 года энергия столкновения частиц на ускорителе достигла  $\geq 130$  ГэВ (LEP-II), что гораздо выше порога рождения  $WW$  пар. Рождение пары  $WW$  бозонов может быть описано тремя диаграммами Фейнмана, показанными на рисунке 1.3.3 Измеренные отношения сечений в лептонной и адронной моде распада измеренной на всех четыре экспериментах LEP представлены на рисунке 1.3.4. Все результаты в хорошем согласии

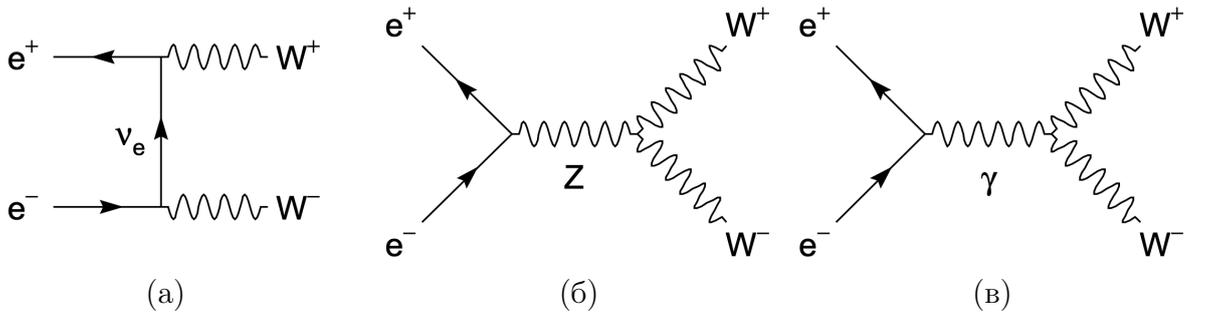


Рисунок 1.3.3 – Диаграммы рождения пар  $WW$  бозонов на ускорителе LEP2. Рождение по  $t$ -каналу с обменом нейтрино (а), рождение через  $s$ -канал с обменом  $Z$  бозона (б) и рождение через  $s$ -канал с обменом виртуальным фотоном (в).

со Стандартной моделью. Были вычислены значения взаимных отношения сечений, что дает качественную проверку лептонной универсальности в распадах  $W$  бозона с точностью около 1 %:  $\mathcal{R}_{\mu/e} = 0.993 \pm 0.019$ ,  $\mathcal{R}_{\tau/e} = 1.063 \pm 0.027$  и  $\mathcal{R}_{\tau/\mu} = 1.070 \pm 0.026$ . Отношения для электронов и мюонов соответствуют ожидаемым, однако, в случае участия в процессе  $\tau$  лептона, отношение заметно выше. Если предположить, что связь для электронов и мюонов одинакова, то для тау лептона наблюдается превышение более чем в  $2.6\sigma$ [1].

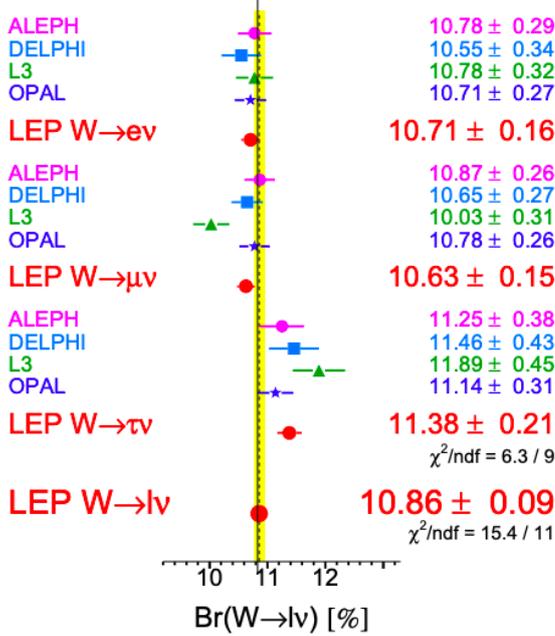
$$2\mathcal{B}(W \rightarrow \tau\bar{\nu}_\tau)/(\mathcal{B}(W \rightarrow e\bar{\nu}_e) + \mathcal{B}(W \rightarrow \mu\bar{\nu}_\mu)) = 1.066 \pm 0.025 \quad (1.3)$$

В тоже время для адронного канала полученный результат находится в поразительном сходстве с предсказаниями Стандартной модели:

$$\mathcal{R}_{q/l} = 1.000 \pm 0.006. \quad (1.4)$$

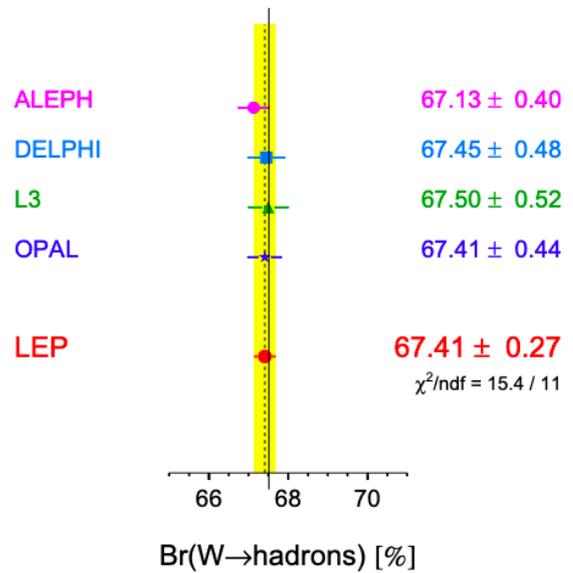
Из последних результатов проверки лептонной универсальности в лептонных распадах  $W$  бозона стоит отметить результат[12] полученный с эксперимента ATLAS[13]. В анализе использовались экспериментальные данные, набранные на детекторе ATLAS в 2015-2018 годах во время второй сессии набора данных с светимостью  $139 \text{ fb}^{-1}$ . При столкновении протон-протонных пучков с суммарной энергией 13 ТэВ. В работе измерялось отношение отношений сечений двух процессов лептонного распада  $W$  бозона  $\mathcal{B}(W \rightarrow \tau\bar{\nu}_\tau)/\mathcal{B}(W \rightarrow \mu\bar{\nu}_\mu)$ , где  $W$ -бозон образовался в результате распада  $t$ -кварка.

### W Leptonic Branching Ratios



(a)

### W Hadronic Branching Ratio



(б)

Рисунок 1.3.4 – Отношение сечений  $W$  бозонов, измеренные на экспериментах ускорителя LEP для лептонного (а) и адронного канала (б).

Статистическая сила измерения обеспечивается большим количеством событий с  $t\bar{t}$  парой, возникающих в протон-протонных столкновениях. Близкое к 100 процентам значение вероятности распада  $t$ -кварка на  $W$  бозон и  $b$ -струю дает большой статистический вклад рождения пар  $WW$  бозонов. В анализе были использованы события с двумя лептонными распадами пары  $WW$  бозонов. Данный критерий применяется для использования техники зондирующего и помечающего лептона. Лептон, обладающий большей энергией, при регистрации помечает событие как интересное и активирует триггер эксперимента на запись. После чего может быть зарегистрирован лептон с более низкой энергией. При проведении измерения отношения отношений сечений использовался только второй лептон. Данная техника позволила регистрировались более мягкие области спектра поперечного импульса  $p_T > 5$  ГэВ лептона. При анализе так же было учтено время жизни промежуточного  $\tau$  лептона, которое влияет на ширину распределения прицельного параметра  $d_0$ . Помимо различий в спектре прицельного параметра  $d_0$  лептонных распадов  $W$  бозона, имеются различия в форме распределения поперечного импульса  $p_T$ . Поэтому переменные  $d_0$  и  $p_T$  были использованы для измерения отношения отношений сечений, которое

составило:

$$\mathcal{R}_{\tau/\mu} = 0.992 \pm 0.013 \quad (1.5)$$

Результат, который был получен в результате данного измерения совпадает с теоретическими предсказаниями Стандартной модели в пределах статистической и систематической погрешностей. На рисунке 1.3.5 представле-

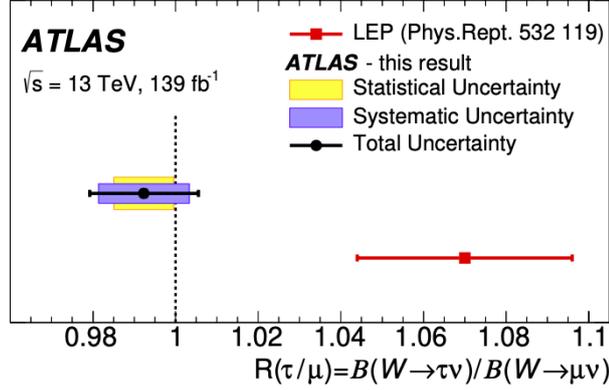


Рисунок 1.3.5 – Экспериментальные результаты по измерению  $\mathcal{R}_{\tau/\mu}$ . Желтым цветом показана статистическая погрешность, синим цветом систематическая погрешность, а черным полная погрешность измерения.

но сравнение значения  $\mathcal{R}_{\tau/\mu}$ . Вертикальная пунктирная линия, установленная по середине, указывает на лептонную универсальность, предсказанную Стандартной моделью. Новое измерение показано черным маркером, а измерение, которое было сделано на LEP красным маркером.

Помимо нового результата по измерению лептонной универсальности с эксперимента ATLAS можно отметить результат, полученный в 2021 году с эксперимента CMS[14], который как и ATLAS проводится на Большом адронном коллайдере. В анализе использовались экспериментальные данные, набранные на детекторе CMS[15] в 2016 году с светимостью  $35.9 \text{ fb}^{-1}$ . Стратегия анализа схожа с той, которую применяли при анализе данных с эксперимента ATLAS, но в анализе данных с CMS не используют переменную прицельного параметра. Отсутствие переменной прицельного параметра, связано с отличием трековых подсистем экспериментов ATLAS и CMS.

В таблице 1.3.1 перечислены значения отношения сечений, полученные на основе описанных выше измерений в экспериментах CMS[14], LEP[1] и ATLAS[12].

	CMS	LEP	ATLAS
$\mathcal{B}(W \rightarrow \mu\bar{\nu}_\mu)/\mathcal{B}(W \rightarrow e\bar{\nu}_e)$	$1.009 \pm 0.009$	$0.993 \pm 0.019$	—
$\mathcal{B}(W \rightarrow \tau\bar{\nu}_\tau)/\mathcal{B}(W \rightarrow e\bar{\nu}_e)$	$0.994 \pm 0.021$	$1.063 \pm 0.027$	—
$\mathcal{B}(W \rightarrow \tau\bar{\nu}_\tau)/\mathcal{B}(W \rightarrow \mu\bar{\nu}_\mu)$	$0.985 \pm 0.020$	$1.070 \pm 0.026$	$0.992 \pm 0.013$
$\mathcal{R}_{\tau/l}$	$1.002 \pm 0.019$	$1.066 \pm 0.025$	—

Таблица 1.3.1 – Отношение отношений сечений лептонных распадов  $W$ , измеренных в анализе данных с CMS[14], LEP[1] и ATLAS[12].

## 1.4 Машинное обучение

С помощью машинного обучения можно выполнять различные задачи, как в физике высоких энергий, так и в других областях науки. В данной работе использовалась TMVA (Toolkit for Multivariate Data Analysis with ROOT) [16] — библиотека алгоритмов машинного обучения, которая идёт в дополнение к пакету анализа больших данных ROOT. На основе TMVA в работе проводилось обучение классификаторов. В этой работе было исследовано влияние параметров классификаторов на их способность классифицировать данные (см. раздел 1.4.3). В исследовании были использованы классификаторы BDT (см. раздел 1.4.2) и kNN (см. раздел 1.4.1). На основе результата, описанного в разделе 3.2 выбирается классификатор и его параметры. Выбранный классификатор используется для классификации событий, прошедших отбор сигнального региона (см. раздел 4.4). Распределение переменной отклика классификатора используется при проведении измерения отношения отношений сечений (см. раздел 5) с использованием метода максимального правдоподобия (см. раздел 1.5.1).

### 1.4.1 Классификатор – метод ближайших соседей (kNN)

Метод ближайших соседей (kNN – k Nearest Neighbors) является одним из самых простых и популярным в задачах классификации. Метод основан на предположении о том, что близким объектам в признаковом пространстве соответствуют похожие метки. Для нового объекта метод предполагает найти ближайшие к нему объекты и построить прогноз по их меткам. В процессе обучения алгоритм запоминает все векторы признаков и соответствующие им метки классов. В данном методе необходимо уметь

определять, насколько объекты близки друг к другу. Это не обязательно евклидово расстояние. Это может быть мера близости объектов по разным критериям. Следовательно, для применения метода kNN в пространстве признаков объектов должна быть введена некоторая метрика (т.е. функция расстояния). Для классификации каждого из объектов тестовой выборки необходимо последовательно выполнить следующие операции:

- Вычислить расстояние до каждого из объектов обучающей выборки.
- Отобрать k объектов обучающей выборки, расстояние до которых минимально.
- Класс классифицируемого объекта — это класс, наиболее часто встречающийся среди k ближайших соседей.

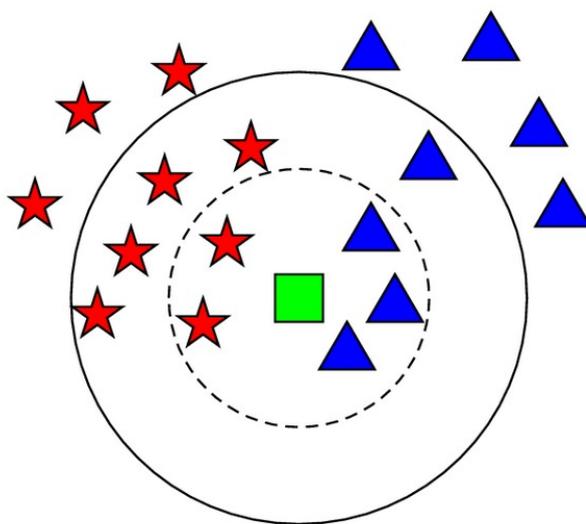


Рисунок 1.4.1 – Иллюстрация работы метода ближайших соседей (kNN)

Согласно методу kNN «зеленый квадрат» (см. рисунок 1.4.1) будет отнесен к классу «синий треугольник» в меньшем круге, а в большем круге к классу «красная звезда». Большое значение имеет число ближайших K - соседей.

## 1.4.2 Классификатор – лес деревьев решений (BDT)

Метод классификации лес деревьев решений (BDT – Boosted Decision Trees) является одним из самых простых и популярных. На этапе обучения

осуществляется отбор данных таким образом, чтобы на выходе получить максимальный прирост информации о данных. На каждом шаге выбирается тот признак, при разделении по которому прирост информации оказывается наибольшим. Далее процедура повторяется рекурсивно.

Дерево решений – это метод представления решающих правил в иерархической структуре, состоящей из элементов двух типов — узлов и листьев. В узлах находятся решающие правила и производится проверка соответствия примеров этому правилу по какому-либо атрибуту обучающего множества (см. рисунок 1.4.2). В простейшем случае, в результате проверки,

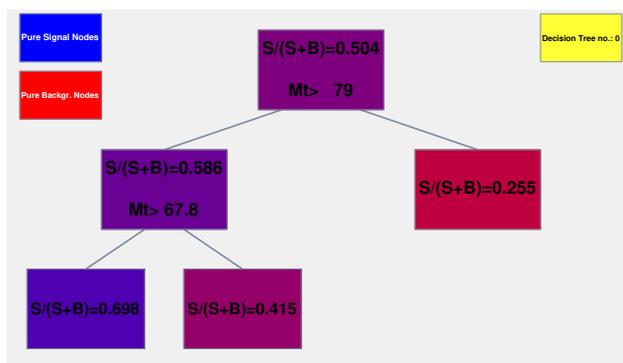


Рисунок 1.4.2 – Схематическое представление дерева решений. Начиная с корневого узла, к данным применяется последовательность разбиений с использованием различающей переменной. Каждое разбиение выбирается таким образом, чтобы обеспечить наилучшее разделение между сигналом (S) и фоном (B).

множество примеров, попавших в узел, разбивается на два подмножества, в одно из которых попадают примеры, удовлетворяющие правилу, а в другое — не удовлетворяющие. Затем к каждому подмножеству вновь применяется правило и процедура рекурсивно повторяется пока не будет достигнуто некоторое условие остановки алгоритма. В результате в последнем узле проверка и разбиение не производится и он объявляется листом.

Лес деревьев решений – это множество решающих деревьев. При построении каждого нового дерева учитывает «опыт» прошлых деревьев таким образом, что бы оно минимизировало ошибку всех предыдущих деревьев. Результатом обучения является дерево, которое имеет минимальную ошибку при классификации. Такой подход называется градиентным бустингом. Важным критерием при построении леса деревьев решений является количество деревьев, а также глубина дерева.

### 1.4.3 Методы оценки классификатора

При классификации данных модель машинного обучения может давать как положительные ответы так и отрицательные. Все ответы классификатора могут быть представлены в виде матрицы ошибок (Confusion Matrix). Матрица ошибок состоит из 4 ячеек (см. таблицу 1.4.1), соответствующим следующим типам откликов классификатора:

- Верно-положительные (TP), объекты, которые были классифицированы как положительные и действительно являются положительными (принадлежащими к данному классу).
- Верно-отрицательные (TN) объекты, которые были классифицированы как отрицательные и действительно отрицательные.
- Ложно-положительные (FP) объекты, которые были классифицированы как положительные, но фактически отрицательные.
- Ложно-отрицательный (FN) объекты, которые были классифицированы как отрицательные, но фактически положительные

	$y = 1$	$y = 0$
$x = 1$	TP	FP
$x = 0$	FN	TN

Таблица 1.4.1 – Матрица ошибок. Где  $x$  - ответ алгоритма на объекте,  $y$  - истинная метка класса на этом объекте

Таким образом ошибки классификатора бывают двух видов: ложно-положительные (FP) или ложно-отрицательные (FN). На основе матрицы ошибок и ее значений, рассчитываются различные метрики классификационной способности алгоритма. В данной работе для оценки качества бинарной классификации используют интеграл под ROC-кривой или кривой ошибок (ROC-integ). Для построения ROC-кривой для каждого классифицируемого объекта вычисляются две метрики оценки. Первой метрикой является "специфичность" (TNR или Background rejection), которая представляет отношение между верно классифицированными негативных событий к

числу всех негативных событий (см. формулу 1.6). Вторая метрика является "полнотой" (TPR или Signal efficiency), которая представляет из себя пропорцию всех верно-положительно предсказанных объектов к общему количеству действительно положительных (см. формулу 1.6). То есть, полнота показывает сколько образцов из всех положительных примеров были классифицированы правильно. Чем выше значение полноты, тем меньше положительных примеров пропущено в классификации.

$$TNR = \frac{TN}{TN + FP}, \quad TPR = \frac{TP}{TP + FN} \quad (1.6)$$

После вычисления данных метрик производится построение их на одном графике (см. рисунок 1.4.3).

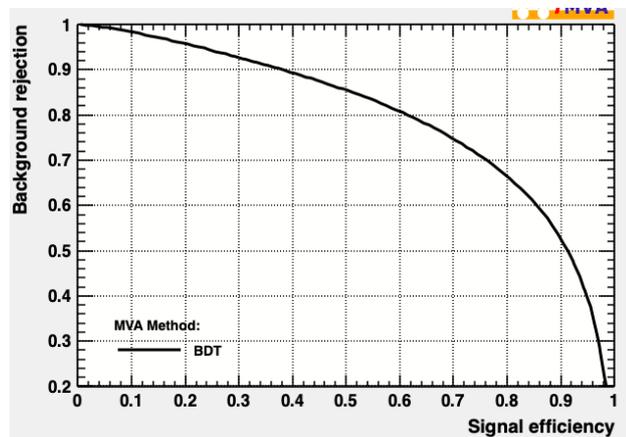


Рисунок 1.4.3 – Пример ROC-кривой. Signal efficiency - TPR (Полнота). Background rejection - TNR (Специфичность)

После чего вычисляется интеграл под ROC-кривой. Чем ближе к единице интеграл под ROC-кривой, тем лучше модель классифицирует события. В этой работе значение интеграла под ROC-кривой используется в исследовании, описанном в разделе 3.2.

## 1.5 Метод максимального правдоподобия

Оценка максимального правдоподобия является популярным статистическим методом. Суть метода максимального правдоподобия заключается в нахождении такого значения оценки на параметр  $\theta$ , при котором значение функции правдоподобия принимает максимальное значение для данной выборки. Так как логарифм функции правдоподобия  $\ln L$  достигает

максимума при тех же значениях искомым параметров, что и сама функция  $L$ , то удобнее максимизировать логарифм функции правдоподобия. Для нахождения интересующих параметров необходимо взять производную от логарифма функции правдоподобия по всем интересующим параметрам и приравнять к нулю,

$$\frac{\partial \ln L(X, \theta)}{\partial \theta_k} = 0, k = 1, 2 \dots m, (\theta = \hat{\theta}) \quad (1.7)$$

составить и решить систему уравнений. Где  $X = x_1, x_2, \dots, x_n$  - фиксированные числа, полученные при измерении случайной величины, а  $\hat{\theta}$  величина является оценкой на параметр.

### 1.5.1 Функция правдоподобия

Для построения функции правдоподобия и дальнейшего фитирования используются гистограммы, которые описывают плотность вероятности событий в заданных регионах селекции. Создание гистограмм производится с помощью `xTauReader` [17]. Вычисление оценки на интересующий параметр производится с помощью `TRExFitter` [18]. Перед началом проведения фитирования необходимо правильно задать функцию правдоподобия и интересующий параметр. Согласно стратегии анализа интересующим параметром в работе является отношение отношений сечений  $\mathcal{R}_{\tau/\mu}$ .

$$\mathcal{R}_{\tau/\mu} = \frac{\mathcal{B}(W \rightarrow \tau\nu)}{\mathcal{B}(W \rightarrow \mu\nu)} \quad (1.8)$$

Покажем, что параметр интереса является отношением отношений сечений. Экспериментальные данные сигнального региона можно выразить через Монте-Карло данные формулой 1.9.

$$N_{data} = \mu_W(\mu_{sig}N_\tau + N_l) + \mu_{ZR}N_{MC} + N_{QCD}, \quad (1.9)$$

В формуле 1.9 переменная  $\mu_{sig}$  является интересующим параметром, а  $\mu_W$  и  $\mu_{ZR}$  нормировочными коэффициентами. Вклад, который вносит распады  $W \rightarrow \tau\nu$  и  $W \rightarrow \mu\nu$  обозначается как  $N_\tau$  и  $N_l$ , а за вклад КХД фона отвечает  $N_{QCD}$ . Доля всех электромагнитных процессов может быть обозначена как  $N_{MC}$ . Использование нормировочных коэффициентов  $\mu_W$  и

$\mu_{ZR}$  оправдано тем, что возможны отклонения нормировок Монте-Карло симуляций и их расхождения с экспериментальными данными. В противном случае, без учета этих нормировочных параметров, параметр интереса может быть оценен неверно. Обозначим отношение отношений сечений распада  $W \rightarrow \tau\nu \rightarrow \mu\nu\nu$  к  $W \rightarrow \mu\nu$  за  $\mathcal{R}_{\tau/\mu}^{exp}$ :

$$\mathcal{R}_{\tau/\mu}^{exp} = \frac{\mathcal{B}(W \rightarrow \tau\nu)\mathcal{B}(\tau \rightarrow \mu\nu\nu)}{\mathcal{B}(W \rightarrow \mu\nu)} \quad (1.10)$$

Тогда, параметр интереса можно задать формулой 1.11. Большинство параметров в процессе генерации Монте-Карло могут быть заданы исходя из результатов предыдущих экспериментов. Например, в этой работе для Монте-Карло было установлено  $\mathcal{B}(\tau \rightarrow \mu\nu\nu)$  равное 17.39 процентам. А с учетом того, что для экспериментальных данных  $\mathcal{B}(\tau \rightarrow \mu\nu\nu)$  соответствует тому же значению, то на эту величину можно сократить числитель и знаменатель. Таким образом получим формулу 1.12.

$$\mu_{sig} = \frac{\left(\frac{\mathcal{B}(W \rightarrow \tau\nu)\mathcal{B}(\tau \rightarrow \mu\nu\nu)}{\mathcal{B}(W \rightarrow \mu\nu)}\right)_{data}}{\left(\frac{\mathcal{B}(W \rightarrow \tau\nu)\mathcal{B}(\tau \rightarrow \mu\nu\nu)}{\mathcal{B}(W \rightarrow \mu\nu)}\right)_{MC}} \quad (1.11)$$

$$\mu_{sig} = \left(\frac{\mathcal{B}(W \rightarrow \tau\nu)}{\mathcal{B}(W \rightarrow \mu\nu)}\right)_{data} / \left(\frac{\mathcal{B}(W \rightarrow \tau\nu)}{\mathcal{B}(W \rightarrow \mu\nu)}\right)_{MC} \quad (1.12)$$

Так как симуляция Монте-Карло основывается на теоретических предсказаниях Стандартной модели, можно приравнять отношение  $\left(\frac{\mathcal{B}(W \rightarrow \tau\nu)}{\mathcal{B}(W \rightarrow \mu\nu)}\right)_{MC}$  к единице. Таким образом получаем формулу 1.13. Из чего можно заключить, что параметр интереса  $\mu_{sig}$  равен  $\mathcal{R}_{\tau/\mu}$ .

$$\mu_{sig} = \mathcal{R}_{\tau/\mu} = \frac{\mathcal{B}(W \rightarrow \tau\nu)}{\mathcal{B}(W \rightarrow \mu\nu)} \quad (1.13)$$

Для проведения измерения используется глобальная функция правдоподобия, которая включает в себя бины распределений и все параметры  $\mu_{sig}$ ,  $\mu_W$  и  $\mu_{ZR}$ , необходимые для проведения измерения. Измерение производится таким образом, чтобы максимизация функции правдоподобия выполнялась по параметрам  $\mu_{sig}$ ,  $\mu_W$  и  $\mu_{MC}$ . При измерении учитываются также систематические погрешности, которые включены в качестве

мешающих параметров  $\theta$ . Мешающие параметры являются параметрами, которые не представляют непосредственного интереса, но которые должны быть учтены при анализе параметров, представляющих интерес. Сама функция правдоподобия определяется стандартным способом в виде произведения распределения Пуассона для каждого бина, каждого используемого региона и функции плотности вероятности  $G$ , которая учитывает систематические погрешности.

$$L = \prod_{b \in \text{ZR}}^{b \in \text{ZR}} \text{Pois}(N_b^{\text{data}}, \mu_{\text{ZR}} N_b^{\text{MC}}) \prod_j^{\text{NPs}} G(\theta_j) \cdot \prod_b^{\text{bins} \in \text{SR}} \text{Pois}(N_b^{\text{data}}, \mu_W [\mu_{\text{sig}} N_b^{\text{T}} + N_b^{\text{I}}] + \mu_{\text{ZR}} N_b^{\text{MC}} + N_b^{\text{QCD}}) \quad (1.14)$$

При задании функции правдоподобия используются три неопределенных параметра:  $\mu_{\text{sig}}$  - интересующий нас параметр, и нормировочные коэффициенты  $\mu_W$  и  $\mu_{\text{ZR}}$ . При получении КХД фона был использован метод оценки КХД фона их экспериментальных данных и интеграл КХД фона заранее определяется с определённой точностью (см. раздел 4.2.3), поэтому нормировочные коэффициент для КХД фона не является неопределённым и не используется в фите. Регионы в которых производится фитирование описаны в разделе 4.2, а описание разбиения на бины этих регионов в главе 5.1.

## 1.6 Стратегия анализа

Сигнатура фонового процесса прямого распада  $W$  в лептон (см. рисунок 1.6.1б) очень схожа с сигатурой сигнального процесса распада  $W$  через  $\tau$ -лептон (см. рисунок 1.6.1а). Для проведения измерения отношения сечений эти процессы необходимо разделить. С помощью методов машинного обучения можно объединить несколько кинематических переменных, за счет чего можно улучшить разделение сигнальных и фоновых процессов.

В разделе 3.2 описан процесс подбора оптимальных параметров для классификаторов BDT (см. раздел 1.4.2) и kNN (см. раздел 1.4.1). Подобра-

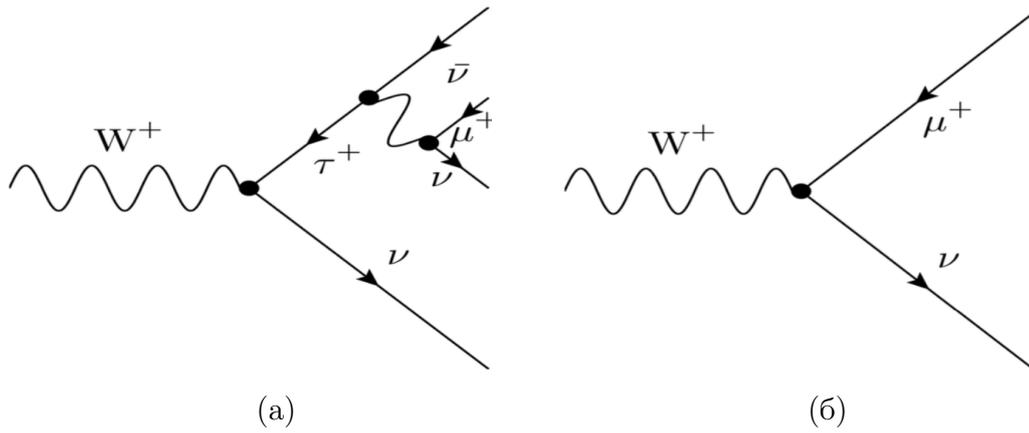


Рисунок 1.6.1 – Диаграммы лептонного распада  $W$  через  $\tau$  лептон (а) и прямого распада  $W$  в  $\mu$  (б)

ны такие параметры, с помощью которых классификатор наилучшим образом будет классифицировать сигнальные и фоновые события. Для оценки качества бинарной классификации в работе используется интеграл под ROC-кривой (см. раздел 1.4.3). Чем больше значение интеграла под ROC-кривой, тем лучше производится классификация фоновых и сигнальных событий. В этом исследовании использовались Монте-Карло данные, соответствующие условиям реальных протон-протонных столкновений эксперимента ATLAS во втором сеансе набора данных в 2015 году в режиме набора данных с большим значением  $\langle \mu \rangle$  (см. раздел 3.1).

В процессе работы, в базовом анализе по измерению лептонной универсальности было принято решение использовать данные, набранные на детекторе ATLAS в режиме с низким значением  $\langle \mu \rangle$ . Поскольку данная работа является частью базового анализа, после проведения работ по выбору оптимального классификатора, в этой работе был совершен переход на данные, набранные на детекторе ATLAS в режиме с низким  $\langle \mu \rangle$  (см. раздел 4.1). При использовании данных с низким  $\langle \mu \rangle$  увеличивается эффективность регистрации частиц, снижается количество КХД фона, который не полностью может быть сгенерирован Монте-Карло генераторами. Помимо снижения фоновых событий становится доступным более низкий порог для триггера на трек лептона. Поэтому далее используется ограничение на поперечный импульс  $p_T > 20$  ГэВ, вместо  $p_T > 25$  ГэВ, как это было с данными с высокой светимостью (см. раздел 3.1.1).

Поскольку был совершен переход на данные с низким значением  $\langle \mu \rangle$ , было заново проведено обучение и валидация классификатора на

новых данных (см. главу 4). При обучении классификатора использовалась модель, полученная с использованием данных с высоким  $\langle \mu \rangle$  (см. раздел 3.2).

Далее в главе 5 описывается процесс получения отношений отношений сечений с использованием новой переменной отклика классификатора. В работе проводится измерение отношения отношений сечений с помощью бинированого метода максимального правдоподобия (см. раздел 1.5.1). Сравниваются результаты измерения с использованием двух разных переменных, а именно переменной отклика классификатора BDT и поперечной массы  $m_T$  (см. формулу 3.1). Получение переменной отклика классификатора BDT описано в главе 5. При проведении фитирования разбиение на бины сигнального региона (см. раздел 4.2.2) производится в двумерном распределенных прицельного параметра  $d_0$  и отклика классификатора BDT или поперечной массы  $m_T$ .

Использование двумерного распределения оправдано тем, что прицельный параметр  $d_0$  (см. раздел 4.2.2) является чувствительной переменной при идентификации распадов  $W$  в  $\tau$ . Исходя из этого можно сделать вывод, что если вместо одномерного распределения отклика классификатора BDT или поперечной массы  $m_T$ , использовать двумерное распределение прицельного параметра  $d_0$  и отклика классификатора BDT или поперечной массы  $m_T$ , то полученный результат отношения отношений сечений будет более точный.

Первоначально измерение необходимо провести в статусе слепого анализа, поэтому вместо экспериментальных данных используются данные, которые полностью соответствуют Монте-Карло симуляции, при этом имеют статистическую погрешность экспериментальных данных. Такие данные называются Азимов данными. Используя данные, которые полностью соответствуют Монте-Карло симуляции можно сделать вывод о том, что полученное значение отношения отношений сечений после фитирования будет равняться единице. В этой работе такой подход возможен, поскольку предметом исследования является не само значение отношения отношений сечений, а его статистическая и систематическая погрешность, которую предполагается снизить за счет использования переменной отклика модели BDT.

## Глава 2

# Экспериментальная установка

В этой главе описывается экспериментальная установка, с использованием данных с которой было проведено исследование. В разделе 2.1 описывается ускорительная установка - Большой адронный коллайдер. В разделе 2.2.1 рассказывается о устройстве и работе детектора ATLAS.

### 2.1 Большой адронный коллайдер

Большой адронный коллайдер (БАК) на сегодняшний день является самым большим и самым мощным ускорителем частиц. Схема его устройства представлена на рисунке 2.1.1. Сам ускоритель является синхротро-



Рисунок 2.1.1 – Схема ускорительного комплекса ЦЕРН

ном, который разгоняет протоны, либо тяжелые ионы по кольцу длиной 26.7 км. В отличие от своего предшественника - ускорителя LEP, БАК имеет более низкие энергетические потери на синхротронное излучение. Благодаря этому при разгоне частиц можно достичь более высоких энергий. При проектировании коллайдера был рассчитан на столкновение протонов с энергией 14 ТэВ и на столкновение ядер свинца с энергией 10 ТэВ в системе центра масс.

Режим работы БАК состоит из нескольких сеансов (RUN), между которыми производится усовершенствование аппаратуры и детекторов, благодаря чему при каждой новой сессии запуска БАК увеличивается энергия столкновения частиц и светимость коллайдера. Во время каждого сеанса работы БАК детекторы набирают данные. До 2021 года на БАКе было 2 сеанса. Первый сеанс набора данных длился с конца 2009 года по 2012 год с энергией столкновения пучков 7 ТэВ и 8 ТэВ. За время первого сеанса наборо-

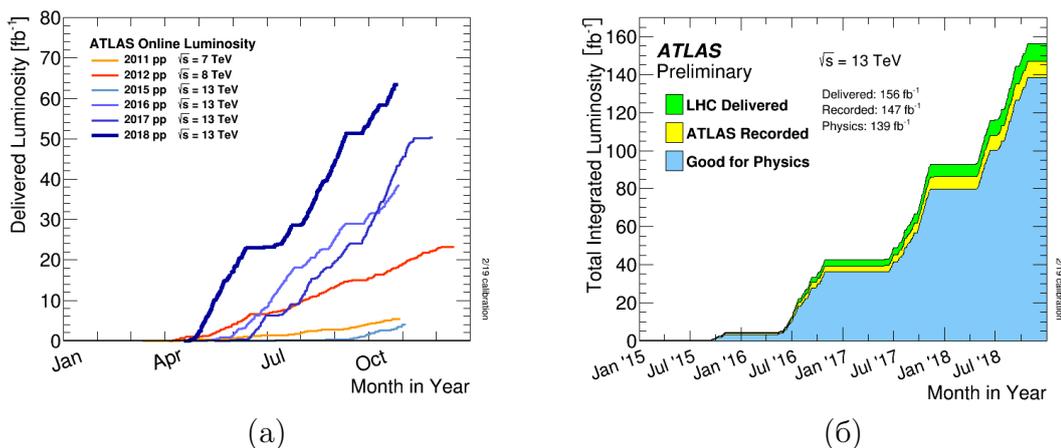


Рисунок 2.1.2 – Интегральная светимость набранная экспериментом ATLAS. Слева показана светимость набранная во время первого сеанса работы (а) в 2011 (оранжевый), в 2012 (красным) и во время второго сеанса работы в 2015-2018(синий.) Справа приведена светимость набранная во время второго сеанса работы (б) в 2015 - 2018 годах, где указана светимость предоставленная ускорителем (зеленый), светимость события непосредственно записанных детектором(жёлтый) и данные с светимостью, сертифицированные как данные хорошего качества (синий) во время стабильных пучков в р-р столкновениях.

ра данных была набрана статистика с интегральной светимостью 30 фб<sup>-1</sup>. Второй сеанс набора данных продолжатся с 2015 года по 2018 год, энергия столкновения составляла 13 ТэВ. За второй сеанс работы БАК была

набрана статистика с интегральной светимостью  $139 \text{ фб}^{-1}$ . В дальнейшем планируется третий сеанс набора данных с энергией  $14 \text{ ТэВ}$ .

В работе были использованы экспериментальные данные, набранные на детекторе ATLAS во время второго сеанса набора данных (см. рисунок 2.1.3). В исследовании, описанном в главе 3 использованы данные, набранные в 2015 году. Далее было решено перейти на данные, которые были

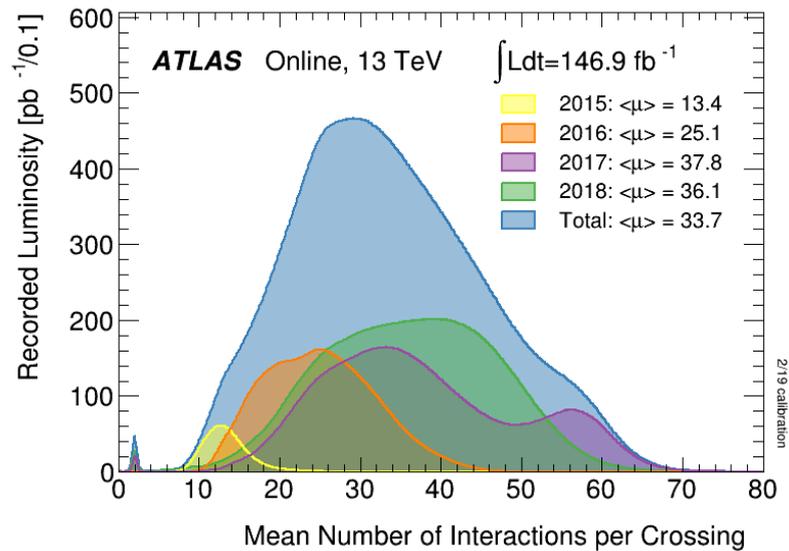


Рисунок 2.1.3 – На рисунке показана набранная светимость на эксперименте ATLAS в зависимости от количества взаимодействий при столкновении. Маленький пик слева отвечает данным, которые были набраны в режиме с низким средним числом взаимодействий  $\langle \mu \rangle$  при столкновении в 2017 и 2018 годах.

набраны на детектора ATLAS в 2017 и 2018 году во время режима набора данных с интегральной светимостью  $340 \text{ пб}^{-1}$ . Поэтому в при обучении и валидации классификатора BDT (см. главу 4), а также при измерении отношения отношений сечений (см. главу 5) были использованы данные с низким средним числом взаимодействий при столкновении  $\langle \mu \rangle$ .

## 2.2 Детектор ATLAS

ATLAS (от англ. A Toroidal LHC ApparatuS) — один из четырёх основных экспериментов на коллайдере LHC в Европейской Организации Ядерных исследований CERN в городе Женева (Швейцария). Детектор ATLAS является комплексной детекторной установкой, которая состоит из

нескольких типов под-детекторов расположенных цилиндрическими слоями вокруг оси трубы ускорителя, симметрично по отношению к направлению движения встречных пучков частиц и с практически полным покрытием по телесному углу вокруг точки столкновения. Схема детектора с обозначением основных элементов показана на рисунке 2.2.1. Детектор

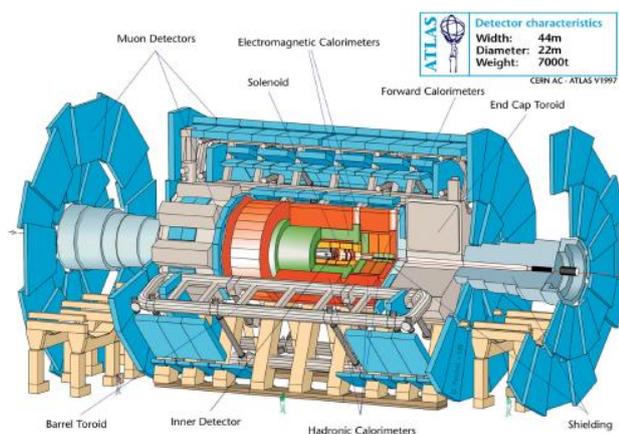


Рисунок 2.2.1 – Детектор ATLAS

состоит из нескольких частей. Для восстановления треков и импульсов заряженных частиц используется внутренний детектор, окруженный сверхпроводящим магнитом, создающим магнитное поле 2 Тл. Заряженные частицы идентифицируются по изгибам их траектории в магнитном поле и таким образом восстанавливается их заряд и импульс. С помощью системы калориметров происходит измерение энергии частиц. Фотоны и электроны высаживают свою энергию в электромагнитном калориметре, в то время как энергия адронов измеряется в адронном калориметре. Мюоны и нейтрино единственные типы частиц пролетающие весь детектор, однако в отличие от мюонов оставляющих треки в мюонном спектрометре, нейтрино пролетают детектор не оставляя следов. Для регистрации мюонов на периферии детектора находится мюонный спектрометр, который предназначен для измерения импульса мюонов. Для отбора событий используется система триггеров. Регистрация нейтрино на детекторе ATLAS не предусмотрена. Поэтому импульс нейтрино может быть представлен как дисбаланс суммарного импульса в поперечной плоскости, а модуль этого вектора может быть представлен как потерянная поперечная энергия.

## 2.2.1 Система координат

Для описания экспериментальных данных с эксперимента ATLAS используется прямоугольная правосторонняя система координат, в которой начало отсчета расположено в номинальной точке столкновения в центре детектора. Ось  $z$  направлена параллельно направлению пучков в трубе ускорителя и плоскость  $x - y$  ему перпендикулярна, причем вектор оси  $x$  смотрит в центр БАК, а вектор оси  $y$  смотрит вверх. Кроме того, используется и цилиндрическая система координат. Полярный угол  $\theta$  отсчитывается от положительного направления оси  $z$ . Азимутальный угол  $\phi$  отсчитывается вокруг оси пучка относительно положительного направления оси  $x$ . Часто используются Лоренц-инвариантные переменные относительно направления пучка, например быстрота:

$$y = \frac{1}{2} \ln \frac{E + p_z}{E - p_z}, \quad (2.1)$$

которая в ультрарелятивистском приближении ( $E \gg m$ ) переходит в псевдобыстроту:

$$\eta = - \ln \tan \frac{\theta}{2}. \quad (2.2)$$

Псевдобыстрота зависит только от полярного угла  $\theta$ . Значение  $\eta = 0$  соответствует поперечной плоскости  $x - y$ , а  $\eta = \pm\infty$  направлению пучка. Угловое расстояние между частицами и треками выражается псевдобыстротно-азимутальным углом:

$$\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}. \quad (2.3)$$

Исходя из определения цилиндрической системы координат, поперечный импульс и энергия равны:

$$p_T = |\vec{p}| \sin \theta = \sqrt{p_x^2 + p_y^2}, \quad (2.4)$$

$$E_T = E \sin \theta = \sqrt{E_x^2 + E_y^2}. \quad (2.5)$$

## Глава 3

# Исследование методов машинного обучения

В данной главе производится подбор оптимального классификатора и его параметров для классификации лептонных распадов  $W$  бозона. В исследовании используется два метода классификации -  $k$  ближайших соседей (см. раздел 1.4.1) и лес деревьев решений (см. раздел 1.4.2).

В разделе 3.1 проведено описание использованных данных в исследованиях. А в разделе 3.2 производится описание оптимизации параметров классификатора.

### 3.1 Используемые данные

Данные используемые в данном исследовании были получены методом Монте–Карло с помощью генераторов PYTHIA для лептонных распадов  $W$  бозона и прошли всю цепочку реконструкций, на условии реальных протон-протонных столкновений эксперимента ATLAS второго сеанса набора данных в 2015 году.

#### 3.1.1 Отбор событий

Для получения региона, который характеризуется большим содержанием сигнальных событий распадов  $W \rightarrow \tau\nu$ , был проведен отбор событий. Кинематический отбор включает в себя ограничение на поперечный импульс лептона  $p_T$  больше 25 ГэВ. Отбор на поперечную потерянную

энергию  $E_T^{miss}$  больше 25 ГэВ, псевдобыстроту  $\eta$  по модулю меньше 2.47, поперечную массу  $m_T$  больше 40 ГэВ, определяемую как

$$m_T = \sqrt{2p_T E_T^{miss}(1 - \cos(\Delta\phi(l - E_T^{miss})))}, \quad (3.1)$$

проводится для отсеечения большого количества КХД фона. Также использовалось ограничение на количество  $b$ -струй, равное 0, на количество лептонов в событии, равное 1. Были применены критерии[19] на идентификацию - medium и изоляцию мюона - tight (см. таблицу 4.2.2 ).

$p_T$	$E_T^{miss}$	$m_T$	$N_{lep}$	$ \eta $	b-струй	ID	ISO
$> 25$ ГэВ	$> 25$ ГэВ	$> 40$ ГэВ	1	$< 2.47$	0	medium	tight

Таблица 3.1.1 – Критерии отбора событий для региона, который характеризуется большим содержанием сигнальных событий распадов  $W \rightarrow \tau\nu$ .

### 3.1.2 Сравнение кинематических распределений мюонного и электронного канала

Перед тренировкой классификаторов было проведено сравнение кинематических распределений для мюонного и электронного канала распадов  $W$  бозона. На рисунках 3.1.1 представлено сравнение кинематических распределений, нормированных на интеграл фоновых процессов, которыми являются прямые распады  $W$  бозона на лептоны. На рисунке 3.1.2 представлено сравнение кинематических распределений, нормированных на интеграл сигнальных процессов. Сигнальным процессом для электронного канала является распад  $W$  через промежуточный  $\tau$  на  $e$  и  $\nu$ . Для мюонного канала сигнальным процессом является распад  $W$  через промежуточный  $\tau$  лептон на  $\mu$  и  $\nu$ . На рисунке 3.1.3 представлено сравнение кинематических распределений, нормированных на интеграл, сигнальных и фоновых процессов. Можно видеть явное отличие сигнального процесса от фонового в поперечной массе, потерянной поперечной энергии, поперечном импульсе лептона.

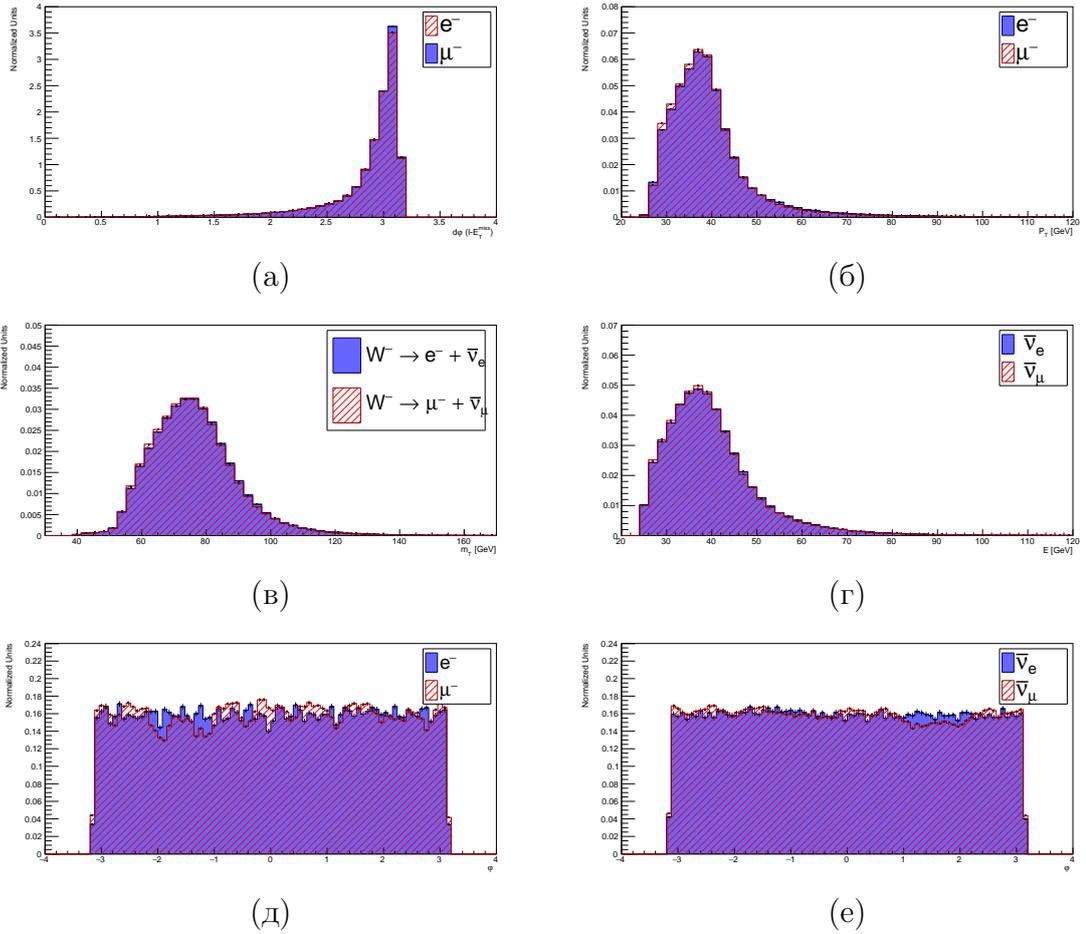


Рисунок 3.1.1 – Сравнение фоновых процессов  $W \rightarrow \mu\nu$  и  $W \rightarrow e\nu$ . Разность азимутальных углов (а) потерянной поперечной энергии и лептона, поперечный импульс лептона (б), поперечная масса (в), потерянная поперечная энергия (г), азимутальный угол лептона (д), азимутальный угол потерянной энергии (е).

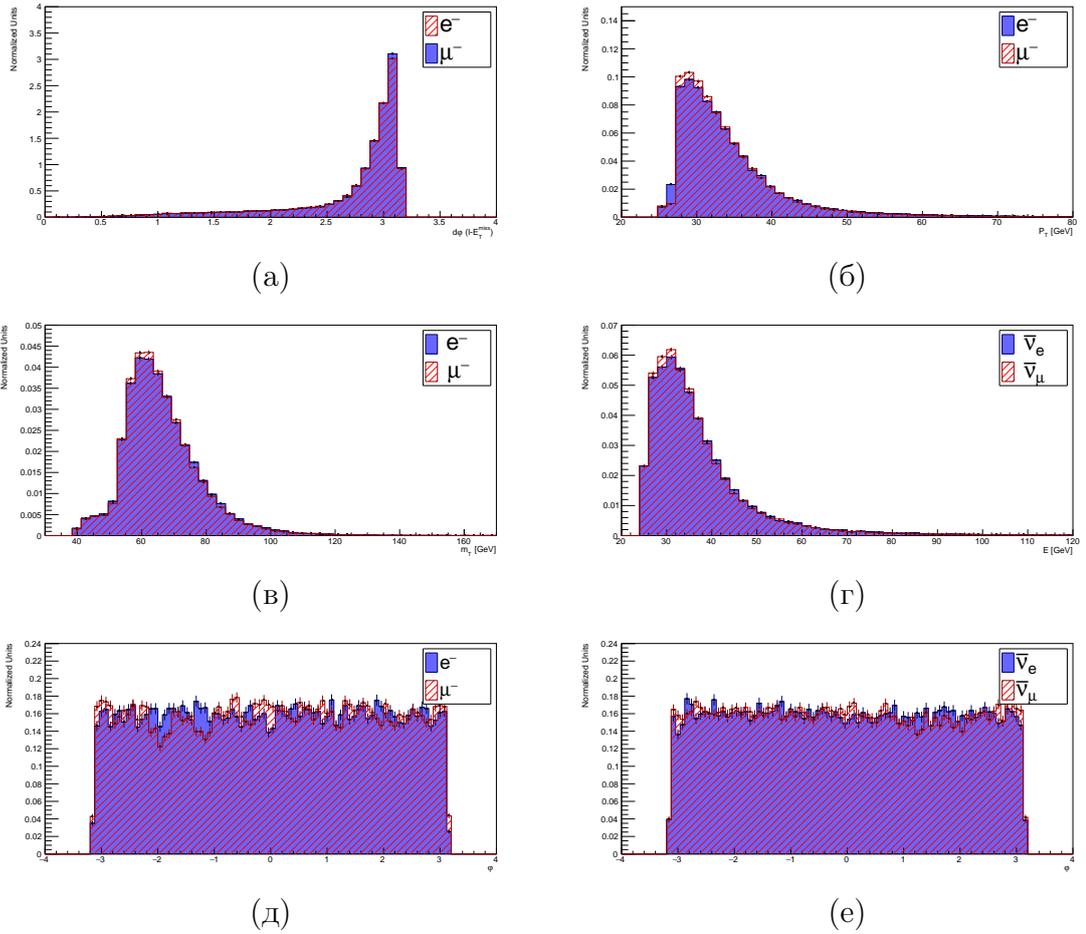


Рисунок 3.1.2 – Сравнение сигнальных  $W \rightarrow \tau\nu \rightarrow e\nu\nu$  и  $W \rightarrow \tau\nu \rightarrow \mu\nu\nu$  процессов. Разность азимутальных углов (а) потерянной поперечной энергии и лептона, поперечный импульс лептона (б), поперечная масса (в), потерянная поперечная энергия (г), азимутальный угол лептона (д), азимутальный угол потерянной энергии (е).

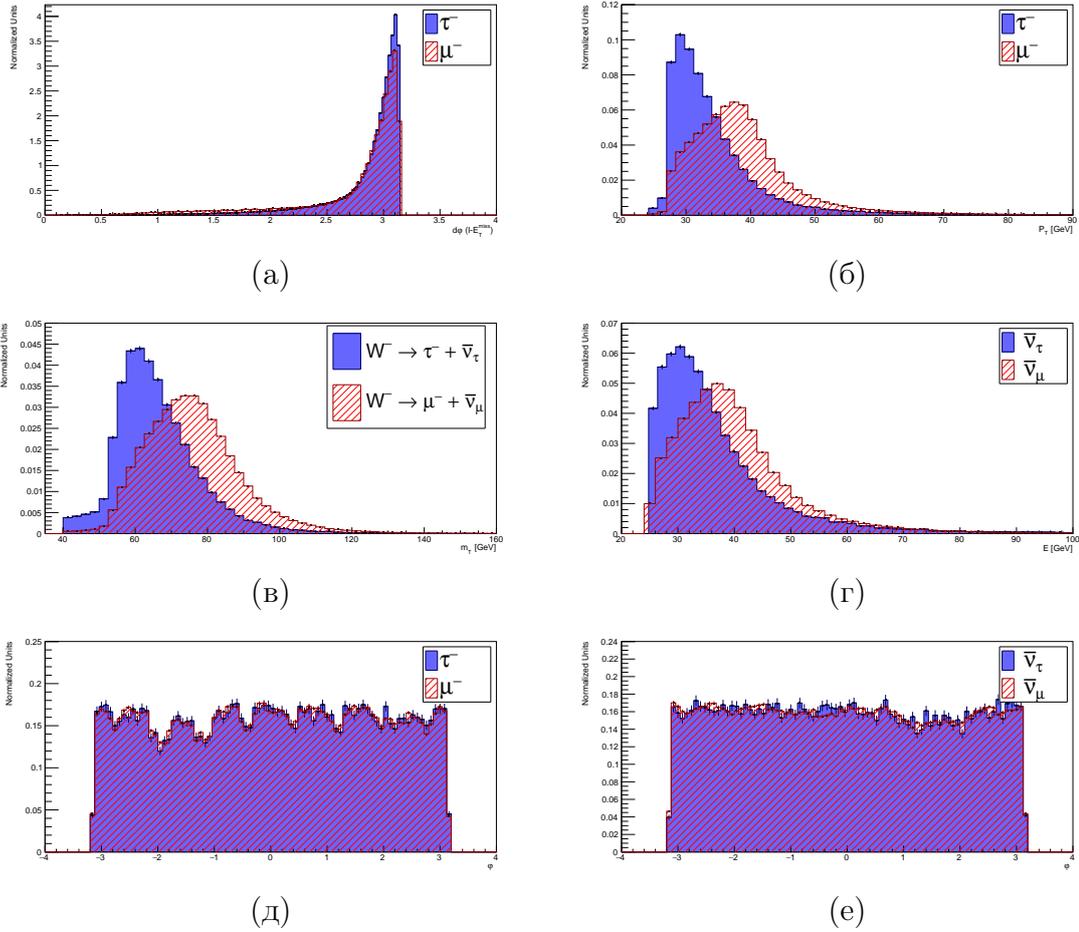


Рисунок 3.1.3 – Сравнение сигнального  $W \rightarrow \tau\nu \rightarrow \mu\nu\nu$  и фонового  $W \rightarrow \mu\nu$  процесса для мюонного канала. Разность азимутальных углов (а) потерянной поперечной энергии и лептона, поперечный импульс лептона (б), поперечная масса (в), потерянная поперечная энергия (г), азимутальный угол лептона (д), азимутальный угол потерянной энергии (е).

Исходя из полученных распределений можно сделать вывод о том, что кинематические распределения для фоновых, как и для сигнальных обоих лептонных каналов одинаковы. Поэтому для экономии затрат вычислительных мощностей далее используется только один канал. Из двух каналов был выбран мюонный, так как он имеет меньше КХД фона чем электронный.

## 3.2 Подбор оптимального классификатора

При оптимизации рассматриваются лептонные распады  $W$  бозона. Как сигнал отмечены распады  $W$  в  $\tau$  с последующим распадом в лептон второго поколения (см. рисунок 1.6.1a), а как фон прямые распады  $W$  (см.

рисунок 1.6.16) в второго поколения.

Для обучения классификатора были использованы переменные:

- Поперечная масса  $m_T$ ;
- Поперечный импульс  $p_T$ ;
- Поперечная потерянная энергия  $E_{\text{miss}}$ ;
- Разность азимутальных углов  $\phi$  для лептона и потерянной энергии  $\Delta\phi(l - E_T^{\text{miss}})$ .

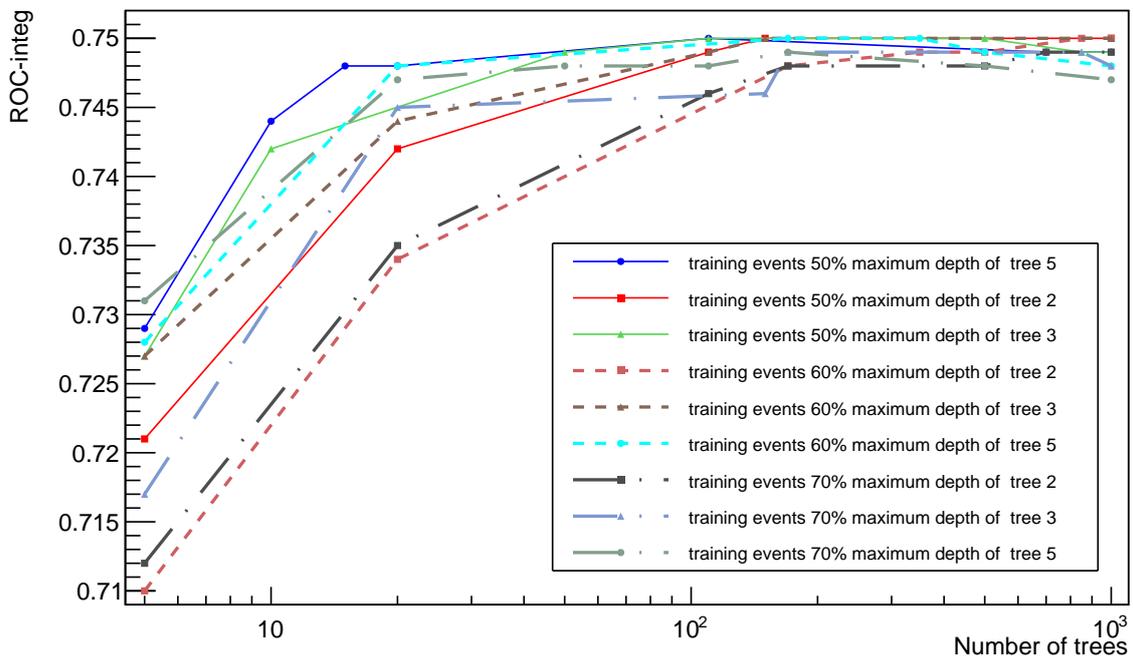
При обучении используется 66779 сигнальных и 496870 фоновых событий. В исследовании рассматриваются отношения тренировочных и тестовых данных: 50%/50%, 60%/40%, 70%/30%. Для BDT варьируются параметры количества деревьев и их глубины, а для kNN количество ближайших соседей.

По данным полученным при варьировании параметров были построены графики зависимости интеграла под ROC-кривой (кривой ошибок) от параметров классификатора BDT и kNN. На (см. рисунок 3.2.1a) представлен график зависимости интеграла под кривой ошибок от количества деревьев. Модели, которые обучались на 50 % данных выходят на плато при значении интеграла под ROC-кривой 0.75. Наилучшими моделями из них являются модели с максимальной глубиной 5 и 3, для них максимум интеграла под кривой ошибок достигается при 110 деревьях. Хороший результат показывает модель с максимальной глубиной 2, для нее максимум достигается при 500 деревьях. Модели, которые обучались на 60 % данных выходят на плато при значении интеграла под ROC-кривой 0.75. Наилучшими моделями являются модели с максимальной глубиной 5 и 3, для максимальной глубины 5 максимум достигается при 170 деревьях, а для максимальной глубины 3 при 150 деревьях. Хороший результат показывает модель с максимальной глубиной 2, для нее максимум достигается при 850 деревьях. Модели которые обучались на 70 % данных тоже показывает хороший результат. Модели с максимальной глубиной 5 и 3 достигают свой максимум интеграла под кривой ошибок 0.749 при 170 деревьях, а при дальнейшем увеличении количества деревьев интеграл под ROC-кривой умень-

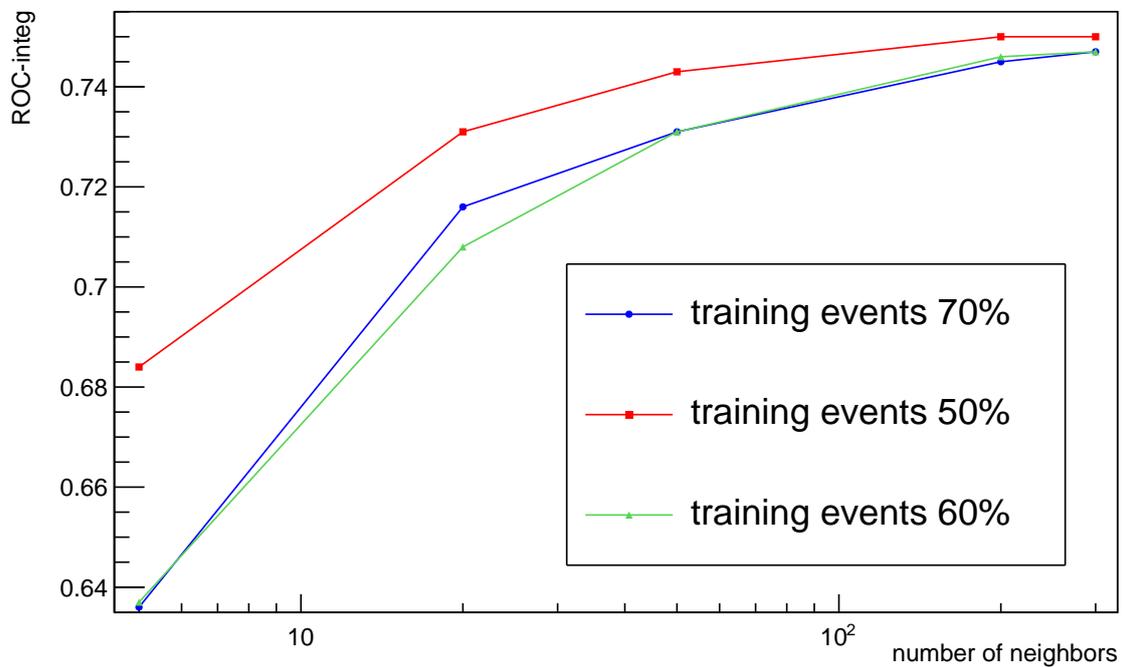
шается. Модель с глубиной деревьев 2 достигает максимум интеграла под кривой ошибок 0.749 при 700 деревьях.

Для классификатора kNN была построена зависимость интеграла под ROC-кривой от количества ближайших соседей (см. рисунок 3.2.16). Для модели которая обучалась на 50% данных максимум интеграла под кривой ошибок 0.749 достигается при 200 ближайших соседях и выходит на плато. Для моделей которые обучались на 60% и на 70% данных максимум интеграла под ROC-кривой 0.747 достигается при 300. Лучший результат показала модель, которая обучалась на 50% данных при 200 ближайших соседях.

Далее в работе для получения отклика классификатора (см. раздел 4.4 ) используется классификатор BDT с 70 % тренировочной выборки, 200 деревьями и глубиной 5.



(a)



(б)

Рисунок 3.2.1 – Зависимость значения интеграла под ROC-кривой от параметров классификатора (а) лес деревьев решений, (б) k-ближайших соседей

## Глава 4

# Исследование отклика классификатора BDT

В этой главе описывается процесс обучения и валидации классификатора BDT. Выбор параметров архитектуры классификатора был проведен в разделе 3.2. В разделе 4.1 описываются данные используемые в работе, а в разделе 4.2 производится описание регионов данных, которые используются при получении значения отношения отношений сечений и при валидации классификатора. После чего в разделе 4.3 описывается процесс настройки и обучения классификатора на данных с низким  $\langle \mu \rangle$ . И, наконец, в разделе 4.4 приведены результаты классификации событий.

### 4.1 Используемые данные

В работе использовались данные с низкой светимостью, полученные с использованием пакета программного обеспечения HistMaker[20].

Экспериментальные данные, используемые в работе, были записаны на детекторе ATLAS в 2017 и 2018 году во время режима набора данных с интегральной светимостью  $340 \text{ пб}^{-1}$  и с низким значением  $\langle \mu \rangle$  (см. раздел 2), в столкновении протон-протонных пучков с суммарной энергией 13 ТэВ.

Смоделированные данные, используемые в работе, были получены методом Монте-Карло с помощью генераторов Pythia[21] и Sherpa[22] и прошли всю цепочку реконструкций, на условии реальных протон-протонных

столкновений эксперимента ATLAS сессия 2 режима с низким  $\langle \mu \rangle$ . Каждому каналу соответствует свой уникальный номер. Список Монте-Карло данных, использованных в работе, приведен в таблице 4.1.1

Процесс	Номер	Генератор	Сечение [пб]
$W^+ \rightarrow \mu\nu$	361101	PowhegPythia8EvtGen	11500.9
$W^- \rightarrow \mu\nu$	361104	PowhegPythia8EvtGen	8579.31
$W^+ \rightarrow \tau\nu$	361102	PowhegPythia8EvtGen	11500.9
$W^- \rightarrow \tau\nu$	361105	PowhegPythia8EvtGen	8579.31
$Z \rightarrow \tau\tau$	361108	PowhegPythia8EvtGen	1950.63
$Z \rightarrow \mu\mu$	361107	PowhegPythia8EvtGen	1950.63
Top	410013	PhPy8EG_P2012	35.8455
Top	410014	PhPy8EG_P2012	35.8244
Top	410470	PhPy8EG	729.77
Top	410642	PhPy8EG	36.993
Top	410643	PhPy8EG	22.174
Top	410644	PowhegPythia8EvtGen	2.06146
Top	410645	PowhegPythia8EvtGen	1.28867
Diboson	363356	Sherpa_221_PDF30	2.20355
Diboson	363358	Sherpa_221_PDF30	3.4328
Diboson	363359	Sherpa_221_PDF30	24.708
Diboson	363360	Sherpa_221_PDF30	24.724
Diboson	363489	Sherpa_221_PDF30	11.42
Diboson	364250	Sherpa_221_PDF30	1.2523
Diboson	364253	Sherpa_221_PDF30	4.579
Diboson	364254	Sherpa_221_PDF30	12.501
Diboson	364255	Sherpa_221_PDF30	3.2344

Таблица 4.1.1 – Список каналов, использованных в генераторе Монте-Карло при моделировании данных с низкой светимостью.

Для сравнения Монте-Карло и реальных данных выполнена нормировка на светимость. Для более точного согласия с распределениями из данных использованы коррекционные коэффициенты, которые учитывают неточности в моделирование Монте-Карло и геометрию детектора. Коэффициенты коррекции, использованные в работе предоставляются Combined Performance (CP) Groups эксперимента ATLAS [23].

## 4.2 Контрольные регионы

В данной работе используется два контрольных региона. Один является сигнальным регионом, а второй является Z регионом. Каждый регион характеризуется отдельным набором ограничений на кинематические переменные и сигнатуру событий.

### 4.2.1 Z регион

Регион распада Z бозона (ZR) используется как контрольный регион, в котором экспериментальные данные очень хорошо согласуются с Монте-Карло данными, а также содержится малое количество КХД фона. С помощью Z региона проводилась валидация сигнального региона, также он использовался для уменьшения систематической погрешности измерения.

Для получения Z региона был проведен отбор событий, который включает в себя ограничение на поперечный импульс  $p_T$  больше 20 ГэВ для двух лептонов. Также применен отбор на инвариантную массу от 66 до 116 ГэВ. Применены критерии на идентификацию и калориметрическую изоляцию (см. раздел 4.2.2).

$p_T$	$M_{ll}$	ID	topoetcone20/ $p_T$
$> 20$ ГэВ	$66$ ГэВ $< m < 116$ ГэВ	tight	$< 0.1$

Таблица 4.2.1 – Критерии отбора событий для Z региона

Для Z региона было выполнено сравнение кинематических распределений в экспериментальных и Монте–Карло данных (см. рисунок 4.2.1). В Z регионе содержится мало КХД фона и Монте–Карло хорошо согласуются с экспериментальными данными. Однако, на отношении в жесткой части поперечного импульса каждого лептона из пары имеются отклонения Монте–Карло от экспериментальных данных. Данное отклонение связано с плохим моделированием Монте-Карло в генераторе PowhegPythia.

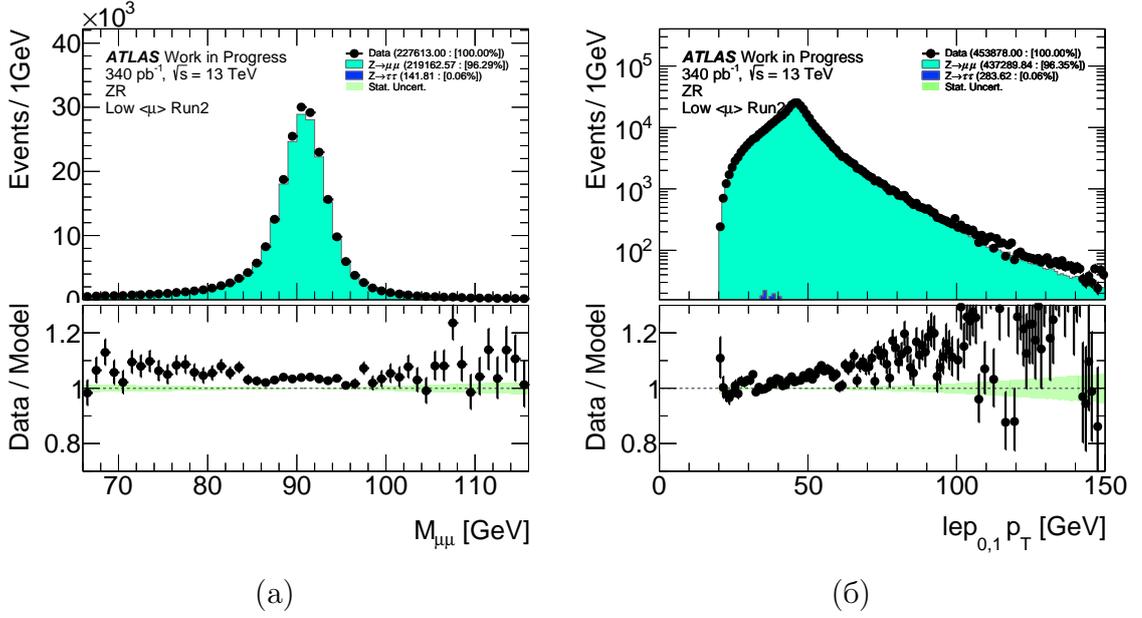


Рисунок 4.2.1 – Сравнение реальных и Монте-Карло данных Z региона для переменных: инвариантная масса (а), поперечный импульс каждого лептона из пары (б)

## 4.2.2 Сигнальный регион

Сигнальный регион (SR) является основным в данной работе и построен так, чтобы максимизировать выход сигнальных событий с распадами  $W$  бозона на промежуточный  $\tau$  лептон. С использованием данных сигнального региона в этой работе был получен отклик классификатора BDT, который используется для измерения отношения сечений (см. раздел 5.3). Для получения сигнального региона был проведен отбор, который включает в себя ограничение на поперечный импульс  $p_T$  больше 20 ГэВ. Отбор на поперечную энергию  $E_T^{miss}$  больше 20 ГэВ, поперечную массу  $m_T$  больше 40 ГэВ, определяемую формулой 3.1. Данный отбор проводится для подавления большого количества КХД фона, который расположен преимущественно в мягкой части спектра. Сигнальным событием является распад  $W$  бозона через промежуточный  $\tau$  лептон, который далее распадается на мюон и нейтрино, не регистрируемое детектором. Поэтому необходимо использовать ограничение на количество лептонов в событии равное 1. Для улучшения выделения лептонов от других, например, не интересующих нас частиц или струй, к лептонам применяется отбор на изоляцию. Трековая изоляция рассчитывается как сумма поперечных импульсов всех треков в конусе размера  $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$ , кроме трека самого

лептона, и делится на импульс лептона. Трековая изоляция характеризует активность вокруг трека в трековом детекторе. Для фоновых частиц активность вокруг трека выше, В работе используется трековая изоляция в конусе  $\Delta R = 0.2$ , равная  $ptvarcone20/p_T < 0.1$ . Калориметрическая изоляция рассчитывается как сумма поперечной энергии в конусе  $\Delta R$ , кроме поперечной энергии самого лептона, разделяется на импульс лептона. Калориметрическая изоляция характеризует активность в калориметре. Для фоновых частиц активность в калориметре вокруг трека выше. В работе используется калориметрическая изоляция в конусе  $\Delta R = 0.2$ , равная  $topoetcone20/p_T < 0.1$ .

$p_T$	$E_T^{miss}$	$m_T$	$N_{lep}$	$ptvarcone20/p_T$	$topoetcone20/P_T$
$> 20$ ГэВ	$> 20$ ГэВ	$> 40$ ГэВ	1	$< 0.1$	$< 0.1$

Таблица 4.2.2 – Критерии отбора событий для сигнального региона

Было выполнено сравнение кинематических распределений для реальных и Монте-Карло данных (см. рисунок 4.2.2 и 4.2.3). Из распределений видно, что большой вклад вносят электромагнитные фоновые процессы. Наибольший вклад 83.35 % вносит процесс прямого распада  $W$  бозона в мюон, значительный вклад 4.82 % вносит также КХД фон и распады  $Z$  бозона на два лептона 3.42 %. Вклад  $Z$  бозона в фоновые процессы происходит из-за того, что иногда один из лептон не удается зарегистрировать, из-за попадания его, например, в пространство, не покрытое детектирующими системами. Поэтому данный распад имеет похожую кинематическую сигнатуру с прямым распадом  $W$  бозона в мюон. На представленных отношениях в кинематических распределениях (см. рисунок 4.2.2 и 4.2.3) отчетливо видно отклонение смоделированных Монте-Карло данных от реальных данных до 20 % в жесткой части поперечного импульса  $p_T$  (см. рисунок 4.2.26), потерянной энергии  $E_T^{miss}$  (см. рисунок 4.2.2г), поперечной массы  $m_T$  (см. рисунок 4.2.2а). Данные отклонения имеются также в  $Z$  регионе, поэтому можно сделать вывод о том, что данное отклонение связано с плохим моделированием Монте-Карло в генераторе PowhegPythia[24].

Переменная прицельного параметра трека  $d_0$  является проекцией расстояния между первичной вершиной взаимодействия и треком. Прицель-

ный параметр  $d_0$  является дополнительным чувствительным показателем к идентификации непрямых распадов  $W$  в  $\tau$ , так как  $\tau$  лептон обладает временем жизни равным  $2.9 \cdot 10^{-13}$  с. Из этого следует, что при распаде  $W$

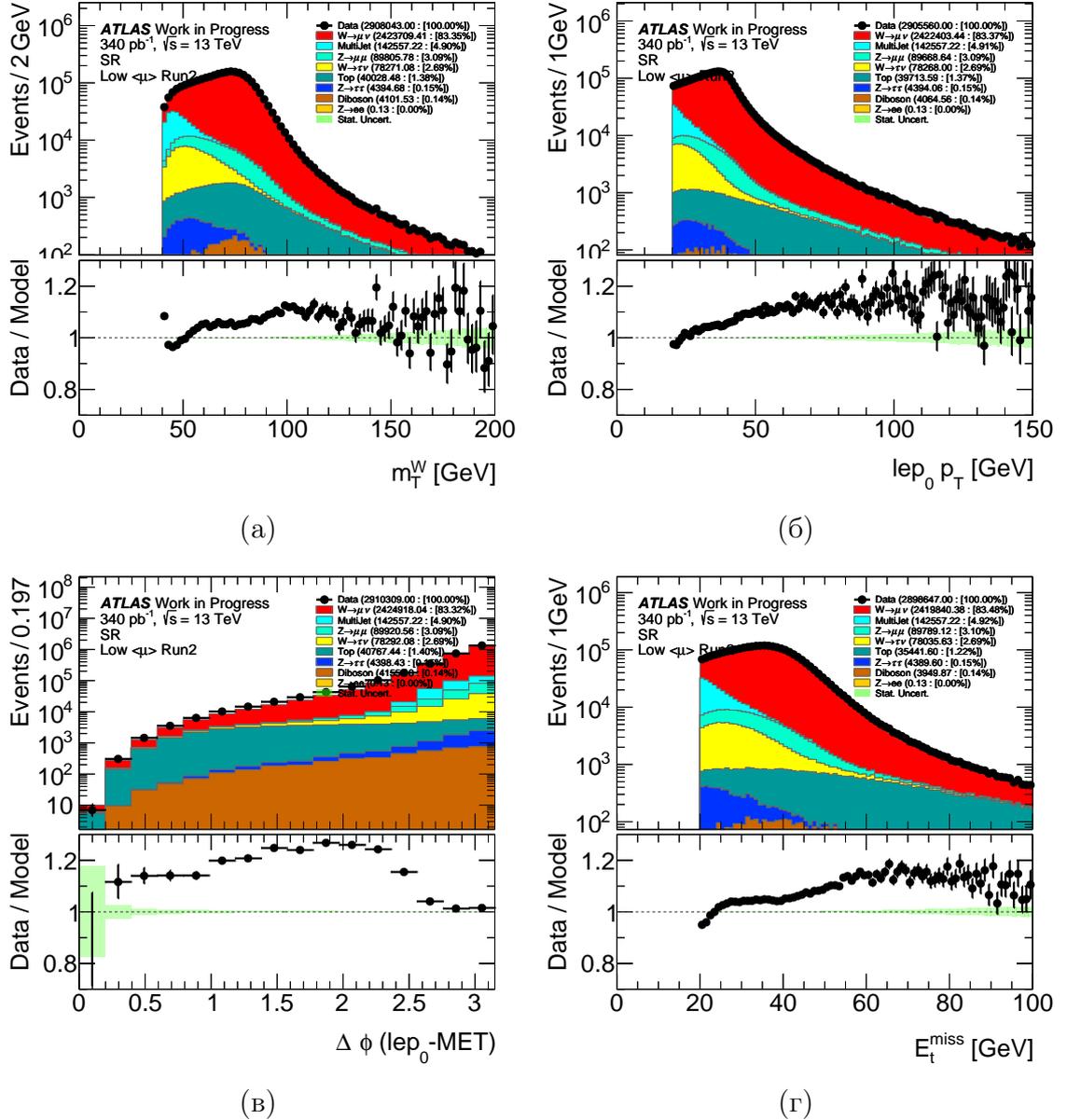


Рисунок 4.2.2 – Сравнение реальных и Монте-Карло данных сигнального региона для переменных: поперечная масса (а), поперечный импульс лептона (б), разность азимутальных углов лептона и потерянной энергии (в), поперечная потерянная энергия (г).

в  $\tau$  распределение прицельного параметра  $d_0$  будет более широкое, чем при прямых распадах  $W$  в  $\mu$  (см. рисунок 4.2.3г). Поскольку  $d_0$  не коррелирует с другими кинематическими переменным, прицельный параметр используется как контрольную переменную при проведении измерения отношения отношений сечений (см. раздел 5.3). Отклонения в отношении эксперимен-

тальных и Монте-Карло данных обусловлены различными значениями сечения пучка протонных сгустков в Монте-Карло и реальных данных.

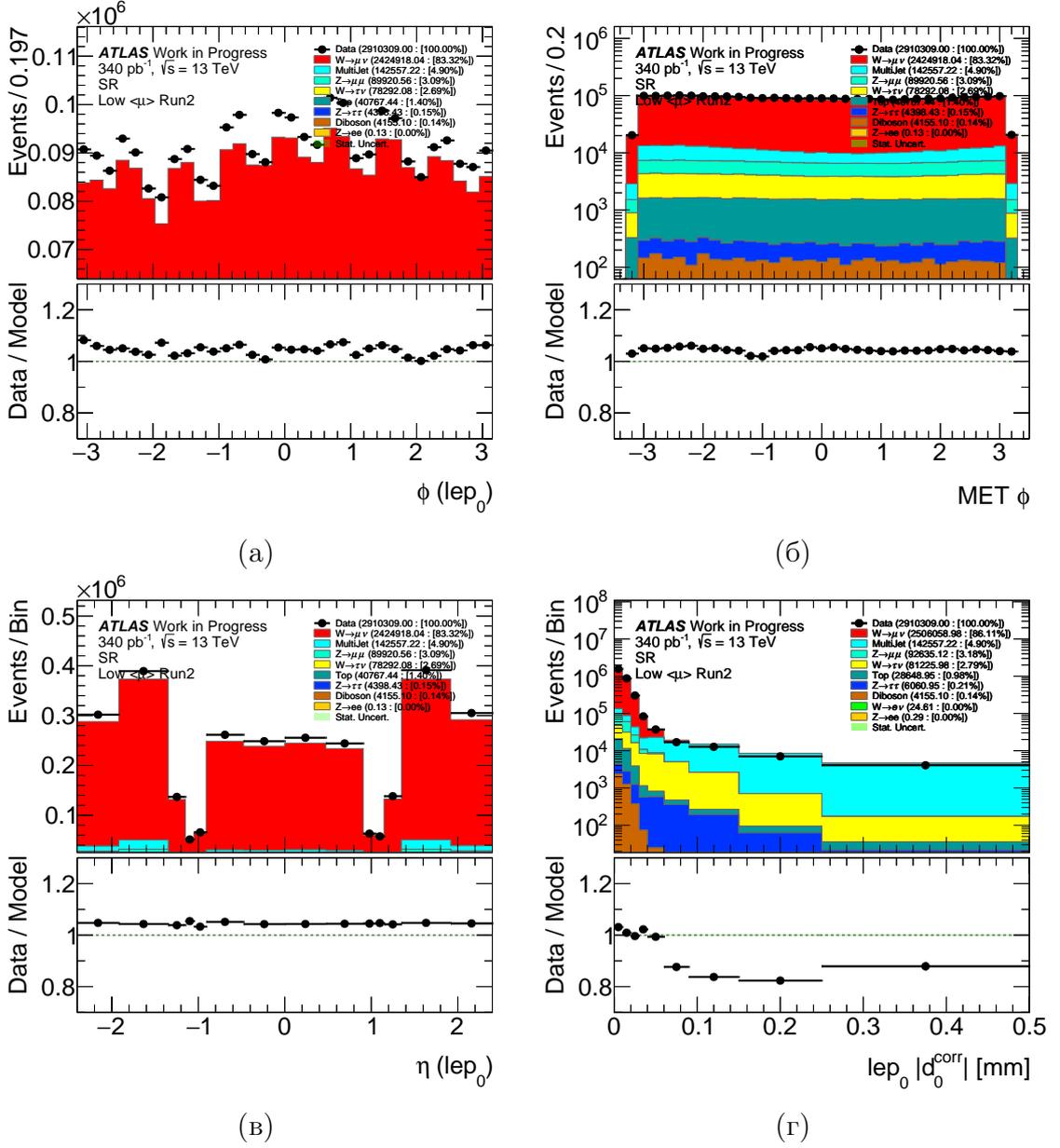


Рисунок 4.2.3 – Сравнение реальных и Монте-Карло данных сигнального региона для переменных: азимутальный угол лептона (а), азимутальный угол потерянной энергии (б), псевдобыстотра лептона (в), поперечный прицельный параметр(г).

### 4.2.3 Оценка КХД фона

Подавление вклада КХД фона в сигнальном регионе достигается за счет отбора изолированных лептонов и ограничений на кинематические переменные:  $E_T^{miss} > 20$  ГэВ,  $p_T > 20$  ГэВ и  $m_T > 40$  ГэВ. Из-за трудно-

стей в точном моделировании процессов, связанных с вкладом КХД, в  $W$  бозонных измерениях используют метод оценки фона из данных[25].

Общий принцип оценки КХД фона из данных строится на выборе региона, обогащенного КХД фоном, путем ослабления и инвертирования одного из выделенных отборов лептонов. Далее определяется несколько КХД шаблонов, разбивающих переменные изоляции лептона для значений превышающих значение используемое в сигнальной области. Далее с использованием шаблонов КХД производится нормировка, с использованием кинематического распределения, с помощью которого можно отделить сигнал от КХД фона. Фитирование выполняется для кинематического распределения в каждом из шаблонов КХД, каждое из которых соответствует определенному срезу по изоляционной переменной. Результатом является построение графика зависимости нормировочного коэффициента от значения изоляционной переменной. Зная эту зависимость можно линейно экстраполировать оценку КХД фона в области сигнального региона. Нормализация вклада КХД фона в области сигнального региона вычисляется путем фитирования шаблонов по трем кинематическим распределениям:  $p_T$ ,  $E_T^{miss}$  или  $m_T$ . КХД фон был посчитан анализ группой[26] и предоставлен в мою работу для использования в исследовании отклика BDT моделей.

#### 4.2.4 Псевдо- $W$ регион

Помимо физических регионов в работе также используется псевдо- $W$  регион, который является искусственно созданным регионом для валидации классификатора BDT. Для создания псевдо- $W$  региона для каждого события из  $ZR$  с вероятностью 50 процентов был выбран один из двух лептонов. После этого выбранный лептон был записан как лидирующий, а второй лептон был добавлен к первоначальной потерянной энергии формула 4.1.

$$E_{T,pseudoW}^{miss} = E_T^{miss} + rand(lep), \quad (4.1)$$

В остальном псевдо- $W$  регион имеет такой же отбор по кинематическим распределениям как и  $Z$  регион. Таким образом мы имитируем сигнатуру сигнального региона, сохраняя при этом хорошее соответствие Монте—Карло и экспериментальных данных. Благодаря такому подходу псевдо-

$W$  регион может быть использован для дополнительной валидации отклика классификатора.

### 4.3 Обучение классификатора BDT

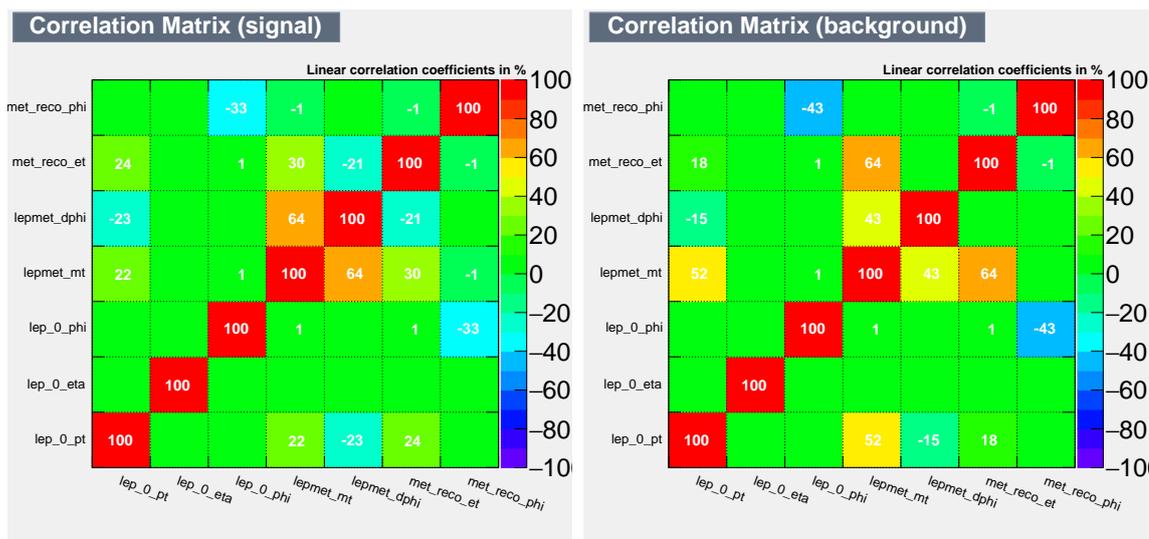
В процессе обучения классификатора BDT был проведен ослабленный предварительный отбор, описанный в главе 4.2.2. Ослабленный предварительный отбор включает в себя кинематический отбор на поперечный импульс  $p_T$  лептона, больше 20, а также количество лептонов, равное одному. Использование ослабленного отбора при обучении классификатора оправдано тем, что при оценке КХД фона для отклика BDT необходимо использовать области, которые имеют ослабленные требования к  $E_T^{miss}$  или  $m_T$ . Ослабление отбора повышает статистику тренировочной выборки.

В качестве сигнала рассматриваются распады  $W$  в  $\tau$  с последующим распадом в лептон второго поколения (см. рисунок 1.6.1a), а в качестве фона рассматриваются прямые распады  $W$  в лептон второго поколения (см. рисунок 1.6.1б), распады  $t$  кварка, дибозонные процессы, распады  $Z$  бозона на лептоны. Полученная выборка для сигнального региона содержит 182806 событий, а в свою очередь полученная выборка для фоновых событий содержит 732000 событий. Соотношение тренировочных и тестовых данных было выбрано 70 на 30 процентов. В обучении модели BDT использовались следующие переменные:

- Поперечная масса  $m_T$ ;
- Поперечный импульс  $p_T$ ;
- Псевдобыстрота лептона  $\eta$ ;
- Азимутальный угол потерянной энергии  $\phi_{miss}$ ;
- Азимутальный угол лептона  $\phi_{lep}$ ;
- Поперечная потерянная энергия  $E_T^{miss}$ ;
- Разность углов  $\phi$  для лептона и потерянной энергии  $\Delta\phi(l - E_T^{miss})$ .

Основываясь на результатах исследования описанного в разделе 4, в котором проводилось исследование зависимости значения интеграла под ROC-кривой от параметров модели BDT, были выбраны параметры модели BDT. Этими параметрами являются количество деревьев - 200 и глубина деревьев - 5.

На рисунке 4.3.1 матрицы корреляции для всех переменных, используемых в обучении. Корреляция поперечной массы  $m_T$  с поперечным им-



(а) Матрица корреляции для сигнала (б) Матрица корреляции для фона

Рисунок 4.3.1 – Матрицы корреляции для сигнала и фона

пульсом  $p_T$  для сигнальных событий составляет 22 процента, а для фоновых 52 процента. В свою очередь, корреляция поперечной массы  $m_T$  с потерянной энергией  $E_T^{miss}$  для сигнальных событий составляет 30 процентов, а для фоновых 64 процентов. Так же имеется корреляция между  $m_T$  и разностью углов лептона и потерянной энергии  $\Delta\phi(l - E_T^{miss})$ . Для сигнала корреляция составляет 64 процента, а для фона 43 процента. Однако, большая корреляция между поперечной массой  $m_T$  и переменными поперечного импульса  $p_T$ , потерянной энергией  $E_T^{miss}$  и разности углов  $\Delta\phi(l - E_T^{miss})$  является вполне ожидаемой (см. формулу 3.1) и не является критичной для классификатора BDT.

Для определения вклада, который вносит каждая переменная при классификации, была посчитана значимость для каждой переменной, использованной при обучении модели BDT (см. таблицу 4.3.1). Наибольший вклад в классификацию вносит поперечная масса  $m_T$ . Это является ожидаемым результатом из-за зависимости поперечной массы  $m_T$  от поперечного

Переменная	$P_T$	$m_T$	$\Delta\phi$	$E_T^{miss}$	$\eta$	$\phi_{miss}$	$\phi_{lep}$
Значимость	1.8e-01	2.6e-01	4.4e-02	1.2e-01	2.3e-04	1.2e-04	1.3e-04

Таблица 4.3.1 – Значимость переменных

импульса  $p_T$  и потерянной поперечной энергии  $E_T^{miss}$ , которые сами вносят значительный вклад при классификации. Наименьший вклад вносят псевдобыстрота лептона  $\eta$ , азимутальный угол для потерянной энергии  $\phi_{miss}$  и лептона  $\phi_{lep}$ .

Значение интеграла под ROC-кривой после тренировки составляет 0.795.

## 4.4 Результат классификации событий

Готовая модель BDT использовалась для классификации данных сигнального региона. На первом этапе классификация проводилась без учета КХД фона. На рисунке 4.4.1a отчетливо видно отклонение Монте-Карло

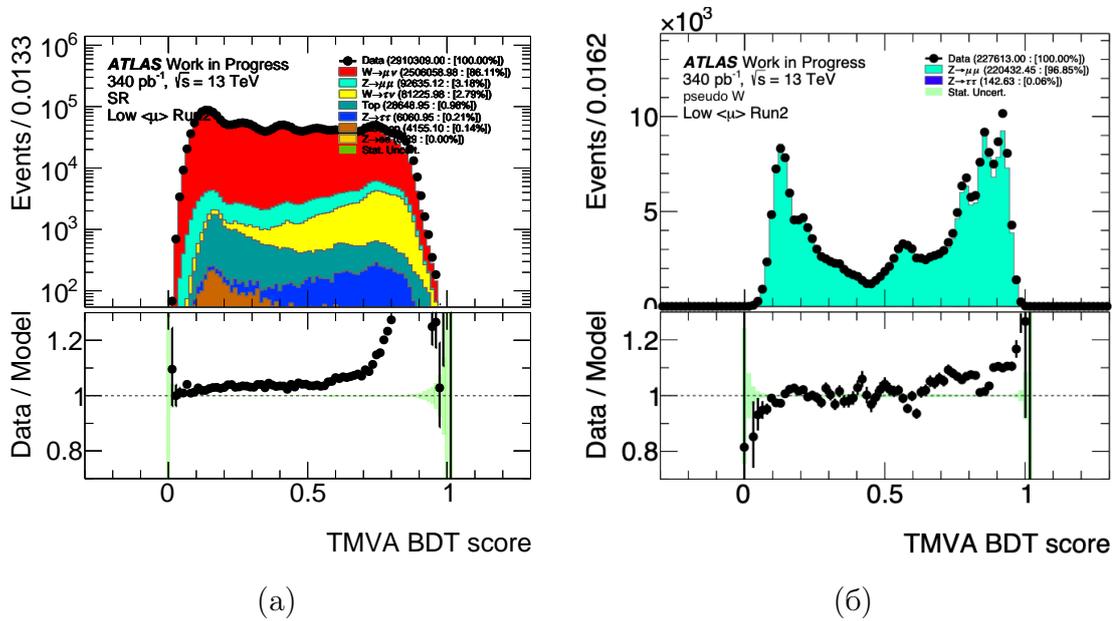


Рисунок 4.4.1 – Слева показан отклик классификатора BDT сигнального региона без КХД фона (а). Справа показан отклик классификатора псевдо-W региона (б).

от реальных данных. Отклонения в области от 0 до 0.6 связаны с плохим моделированием данных Монте-Карло, уже отмеченным ранее. Имеются

также отклонения в области от 0.6 до 1. Предполагается, что они объясняются неучтенным КХД фоном. Для подтверждения данной гипотезы проводится классификация данных псевдо- $W$  региона.

По распределению отклика классификатора в псевдо- $W$  регионе (см. рисунок 4.4.1б) можно сделать вывод, что классификация Монте-Карло и экспериментальных данных хорошо согласуется между собой, а это подтверждает гипотезу, что причиной отклонения в классификации данных сигнального региона в области 0.6 до 1.0 является неучтенный КХД фон, а не расходимость в отношении Монте-Карло и экспериментальных данных. На втором этапе, после как оценка для КХД была посчитана и предоставлена в данный анализ, осуществлялась классификация уже с учетом КХД фона. На рисунке 4.4.2а представлен отклик модели на данные сигнального региона вместе с КХД фоном. Имеются отклонения Монте-Карло от реальный данных в области от 0.9 до 1.0. Причиной отклонений может являться как плохое моделирование Монте-Карло в КХД регионе, так и неполная оценка КХД фона.

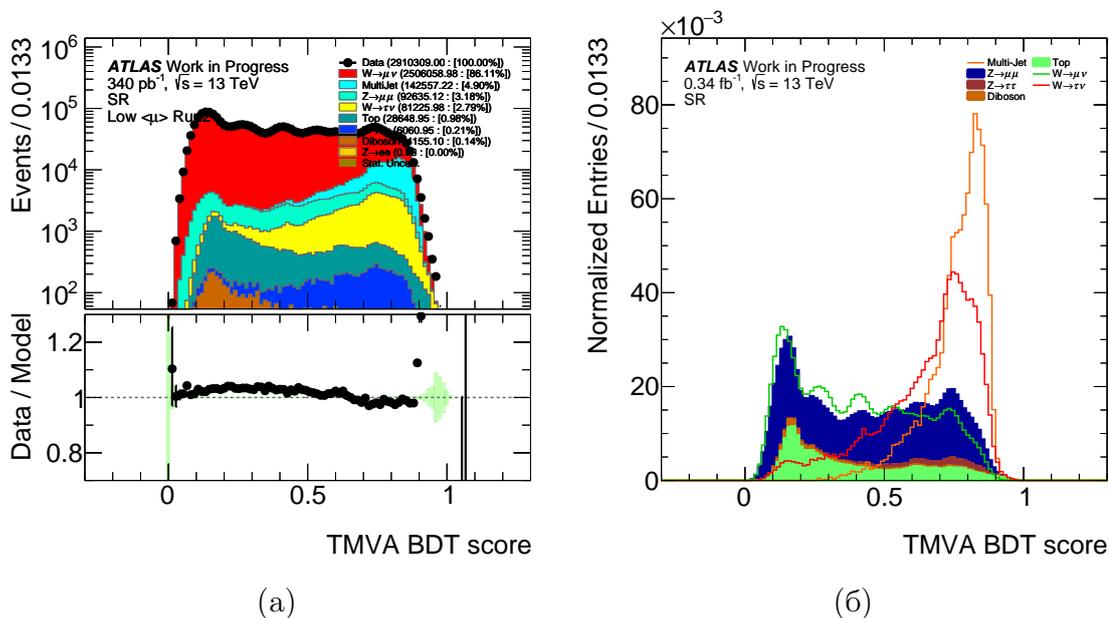


Рисунок 4.4.2 – Слева показан отклик классификатора BDT на данные сигнального региона с КХД фоном (а). Справа показано сравнение отклика классификатора BDT сигнального региона после нормировки на интеграл (б).

Были построены нормированные на интеграл распределения отклика для каждого канала Монте-Карло данных (см. рисунок 4.4.2б). На распределениях видно явное отличие формы сигнала от фона. Форма распреде-

лений отклика для сигнала и КХД является схожей. Это происходит из-за того, что эти процессы имеют схожую сигнатуру, например, в распределении поперечного импульса  $p_T$  (см. рисунок 4.2.26) эти процессы находятся в мягкой части. То же самое можно сказать и о схожести всех фоновых событий, кроме КХД фона. На основе данного исследования можно сделать вывод о том, что модель BDT хорошо справляется с задачей классификации непрямых лептонных распадов  $W$  бозона и электрослабых процессов. Далее в главе 5 описан процесс получения значения отношения отношений сечений с использованием отклика классификатора BDT.

## Глава 5

# Измерение лептонной универсальности

В этой главе описана заключительная часть работы, которой является исследование влияния на точность измерения отношения отношений сечений от использования отклика классификатора BDT. Измерение проводится с использованием бинированного метода максимального правдоподобия, описанного в разделе 1.5.

В разделе 5.1 описан процесс оптимизации бинирования распределения отклика классификатора BDT. Описание систематических погрешностей осуществляется в разделе 5.2. В разделе 5.2 представлен результат измерения отношения отношений сечений с использованием отклика классификатора BDT, а также выполняется его сравнение с результатом измерения с использованием поперечной массы  $m_T$ .

### 5.1 Оптимизация бинирования BDT

Перед началом проведения фитирования была произведена оптимизация бининга для переменной отклика модели BDT. Оптимизация бининга проводится таким образом, чтобы минимизировать систематическую и статистическую погрешность для параметра интереса  $\mu_{sig}$ . Двухмерный фит переменных отклика BDT и прицельного параметра  $d_0$  выполняется путем разбиения событий на 5 областей для отклика BDT, каждая из которых имеет соответствующее одномерное распределение  $d_0$ , которое в свою оче-

редь разбито на 8 бинов. Также производится разбиение инвариантной массы  $Z$  региона на 50 бинов. Таким образом общее количество бинов, которые участвуют в фитировании составляет  $8 \times 5 + 50 = 90$ .

Разбиение распределения BDT на области выбрано таким образом, чтобы получить максимальную чувствительность к изменению формы сигнального и фонового распределения. При выборе разбиения учитывается также отношение количества сигнальных событий к фоновым. На рис. 5.1.1 продемонстрировано разбиение переменной отклика классификатора BDT с учетом формы распределения и отношения сигнальных событий к фоновым. Синими линиями обозначены разбиения распределения отклика классификатора BDT на бины.

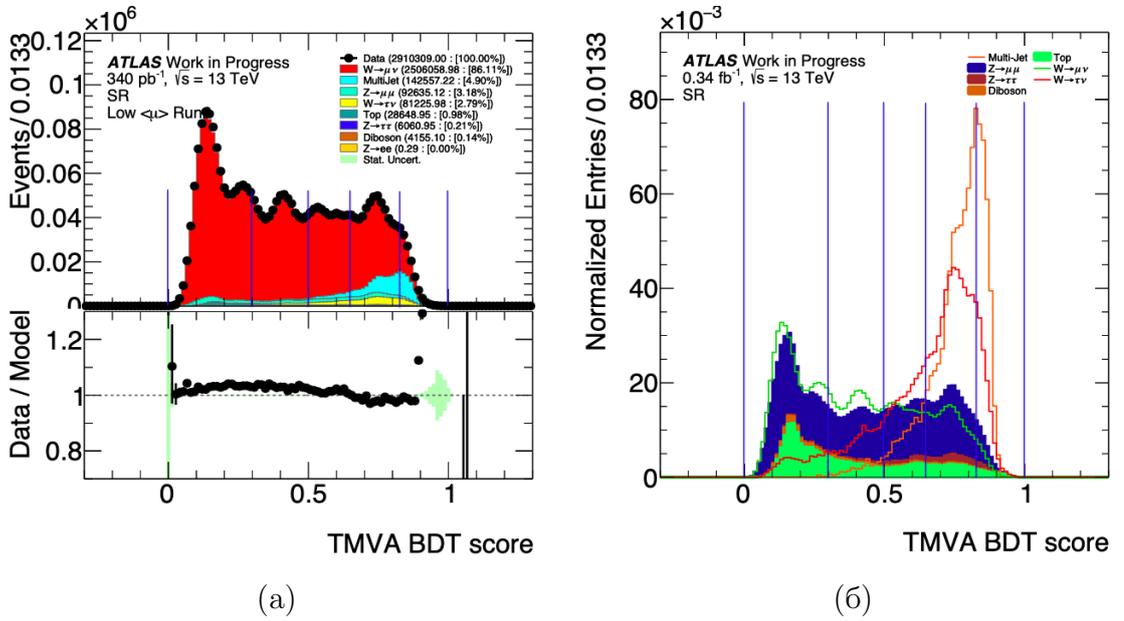


Рисунок 5.1.1 – (а) - Распределения отклика классификатора BDT, (б) - Нормированные распределения отклика классификатора BDT

Оптимизация бинирования привела к следующему разбиению для отклика BDT:  $[0.0, 0.3, 0.5, 0.65, 0.83, 1.0]$ .

Разбиение для  $d_0$  было взято из аналогичного анализа  $t\bar{t}$ -группы [27] эксперимента ATLAS. В работе проводилось исследование зависимости значения отношения сечений от используемой переменной отклика классификатора BDT или поперечной массы  $m_T$  в двухмерном распределении, при постоянном разбиении на бины распределения  $d_0$ , поэтому в работе оптимизация бининга для  $d_0$  не проводилась.

## 5.2 Учет систематических погрешностей

Большинство систематических погрешностей оцениваются в диапазоне  $\pm 1\sigma$ , некоторые из них представляются инструментами анализа эксперимента ATLAS. Систематические погрешности могут быть разделены на три основные категории: экспериментальные, систематические погрешности связанные с оценкой фоновых процессов и теоретические. Теоретические систематические погрешности не были использованы в работе. А систематические экспериментальные и фоновое систематические погрешности в данной работе используются при измерении отношении отношений сечений. Как было указано в разделе 1.5.1 систематические погрешности входят в функцию правдоподобия как мешающие параметры. Для каждого объекта анализа соответствует свой набор параметров систематической погрешности.

Для мюонов рассматривался следующий вклад систематических погрешностей:

- Scale - отвечает за систематическую погрешность связанную с неверным измерением импульса;
- Efficiency - рассматривается вклад в систематическую ошибку от эффективности для эффективности следующих четырех компонент: эффективность триггера, эффективность идентификации, эффективность изоляции и эффективность построения связи с вершинами взаимодействия;
- Reconstruction - отвечает за погрешность на реконструкцию мюона.

Для поперечной потерянной энергии:

- Scale - погрешность масштабирования, рассчитанная с использованием всех составляющих масштабных факторов.

Вклад систематических погрешностей связанных с набором данных на ускорителе учитывался по следующим параметрам:

- Светимость - систематическая погрешность связанная с ошибкой в измерении записанной светимости данных. Это погрешность зафиксирована на уровне 1.5%.

Для оценки КХД фона:

- MJ yield extrapolation - погрешность на интегральный выход КХД фона, коррелирующая от бина к бину;
- MJ template - погрешность на форму распределения КХД фона, которая не является коррелирующей от бина к бину.

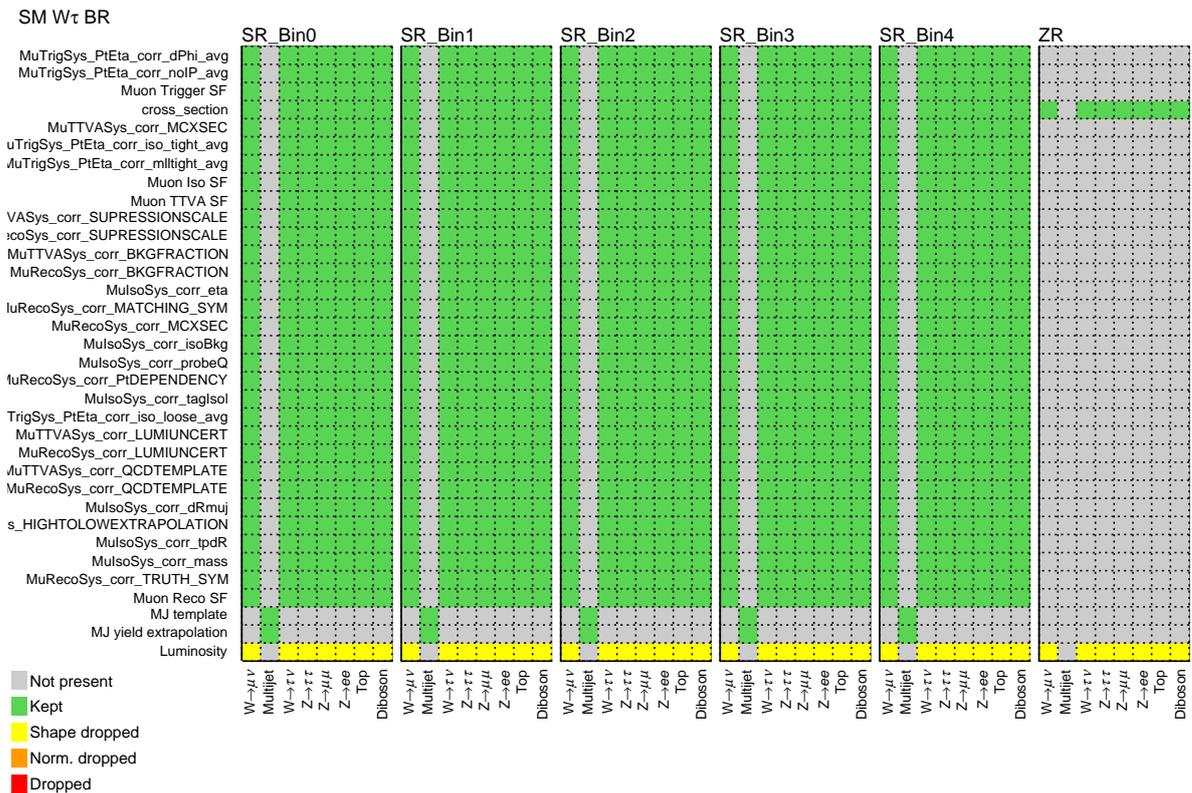


Рисунок 5.2.1 – Используемые систематические погрешности для каждого процесса и региона.

В таблице 5.2.1 представлен список систематических погрешностей, используемых при измерении отношения отношений сечений.

На рисунке 5.2.1 показано в каких регионах и для каких процессов используются те или иные систематические погрешности. Видно, что для Z региона для всех процессов используется только систематическая погрешность на сечение. Оценка КХД фона не используется в Z регионе, поэтому к ней не применяются никакие погрешности в Z регионе. Для оценки КХД фона в сигнальном регионе используется только две систематические погрешности. Для остальных процессов в сигнальном регионе применяются все систематические погрешности, описанные выше, кроме систематических погрешностей оценки КХД фона.

---



---

muon_eff_reco
MuRecoSys_corr_BKGFRACTION
MuRecoSys_corr_HIGHTOLOWEXTRAPOLATION
MuRecoSys_corr_LUMIUNCERT
MuRecoSys_corr_MATCHING_SYM
MuRecoSys_corr_MCXSEC
MuRecoSys_corr_PtDEPENDENCY
MuRecoSys_corr_QCDTEMPLATE
MuRecoSys_corr_SUPRESSIONSCALE
MuRecoSys_corr_TRUTH_SYM
MuTTVASys_corr_BKGFRACTION
MuTTVASys_corr_LUMIUNCERT
MuTTVASys_corr_MCXSEC
MuTTVASys_corr_QCDTEMPLATE
MuTTVASys_corr_SUPRESSIONSCALE
muon_eff_iso
MuIsoSys_corr_dRmuj
MuIsoSys_corr_eta
MuIsoSys_corr_isoBkg
MuIsoSys_corr_mass
MuIsoSys_corr_probeQ
MuIsoSys_corr_tagIsol
MuIsoSys_corr_tpdR
muon_eff_trig
MuTrigSys_PtEta_corr_dPhi_avg
MuTrigSys_PtEta_corr_iso_tight_avg
MuTrigSys_PtEta_corr_noIP_avg
MuTrigSys_PtEta_corr_iso_loose_avg
MuTrigSys_PtEta_corr_mlltight_avg

---



---

Таблица 5.2.1 – Список экспериментальных систематических погрешностей, используемых в работе. Погрешности предоставляются Combined Performance (CP) Groups эксперимента ATLAS[23]

## 5.3 Результаты

В работе проводится исследование влияния используемой контрольной переменной на статистическую погрешность во время двухмерного фитирования. Сравниваются результаты для двух переменных: поперечной массы  $m_T$  и отклика модели BDT. Предполагается, что использование BDT отклика в фитировании обеспечит более точное измерение отношения отношений сечений, так как BDT формируется на основе 7 переменных, одной из которых является поперечная масса  $m_T$ . Фитирование проводится на Азимов данных, которые используются для проверки работоспособности фита, а также для выявления различных проблем, таких как сильно ограниченные мешающие параметры или большие корреляции между параметрами фита. Азимов данные также используются для определения того, какие погрешности оказывают наибольшее влияние на параметр интереса  $\mu_{sig}$  до того как будут использованы реальные данные.

На пером этапе работы было произведено фитирование с использованием поперечной массы. Разбиение переменной поперечной массы  $m_T$  не было оптимизировано в работе, а было взято из базового анализа. На рисунке 5.3.1a продемонстрировано разбиение переменной поперечной массы  $m_T$  на бины, а на рисунках 5.3.2 показано распределение переменной  $d_0$  в каждом бине  $m_T$ . После проведения фитирования было получено значение

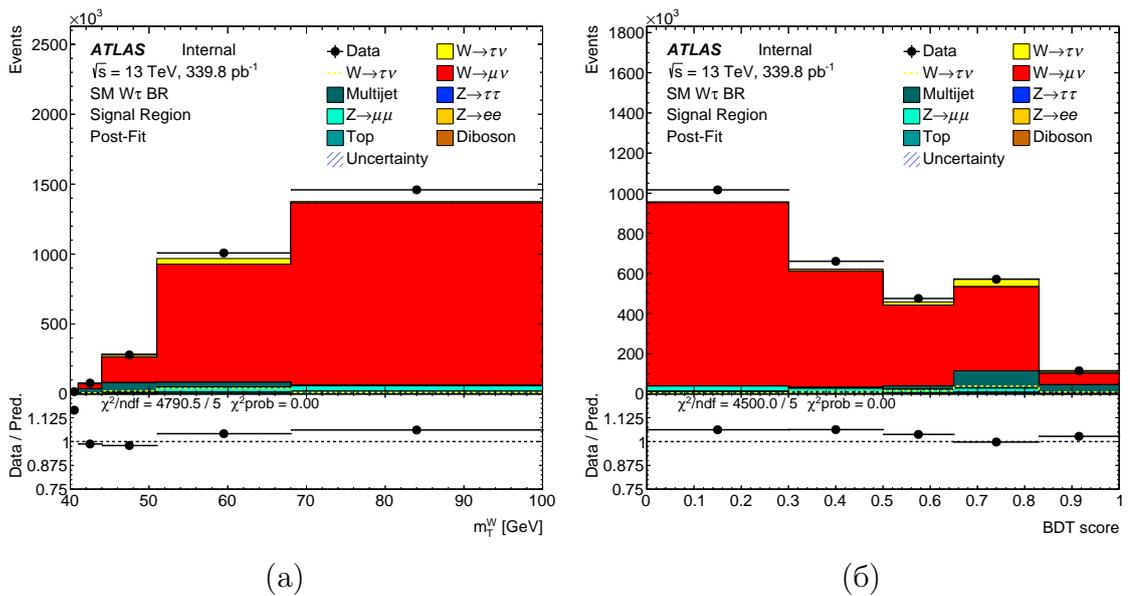


Рисунок 5.3.1 – (а) - Распределения переменной поперечной массы  $m_T$ , (б) - распределения переменной отклика модели BDT.

параметра интереса  $\mu_{sig}$  и его статистической и систематической погрешности:

$$\mu_{sig} = 1.0000_{-0.015}^{+0.015}(stat)_{-0.022}^{+0.022}(syst) \quad (5.1)$$

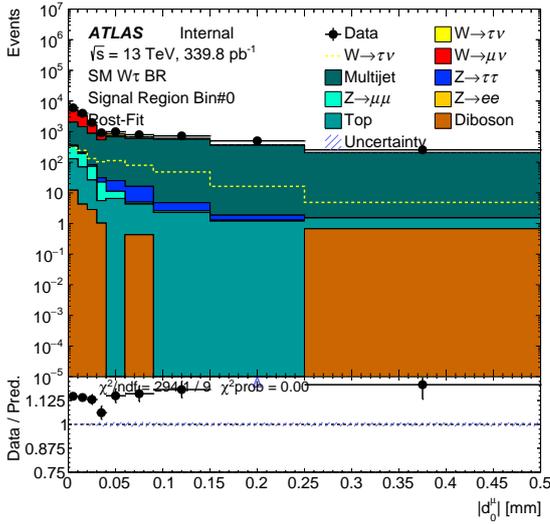
На втором этапе по аналогии с первым было проведено фитирование с использованием новой переменной отклика BDT. На рисунке 5.3.16 показано разбиение для переменной отклика модели BDT, а на рисунке 5.3.3 показано распределение переменной  $d_0$  в каждом бине BDT.

Получено значение параметра интереса и его статистической погрешности:

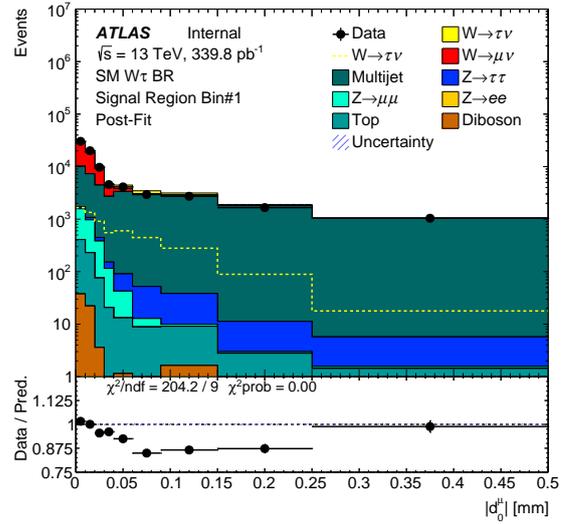
$$\mu_{sig} = 1.000_{-0.015}^{+0.015}(stat)_{-0.027}^{+0.027}(syst) \quad (5.2)$$

Сравнение двух результатов фитирования говорит о том, что использование переменной отклика классификатора BDT при фитировании не влияет на статистическую погрешность данных, при этом увеличивает систематическую погрешность измерения на 0.005.

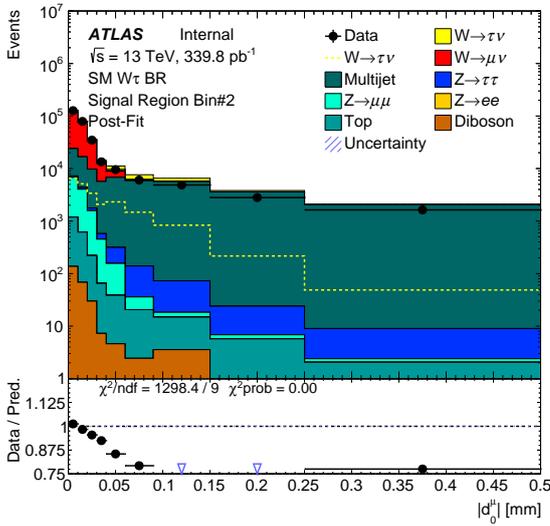
Не смотря на статус слепого анализа, на рисунках 5.3.2 и 5.3.3 показано сравнение реальных и Монте-Карло данных. Это сделано для того, чтобы показать имеющиеся отклонения в распределении  $d_0$  между реальными и Монте-Карло данными (см. раздел 4.2.2).



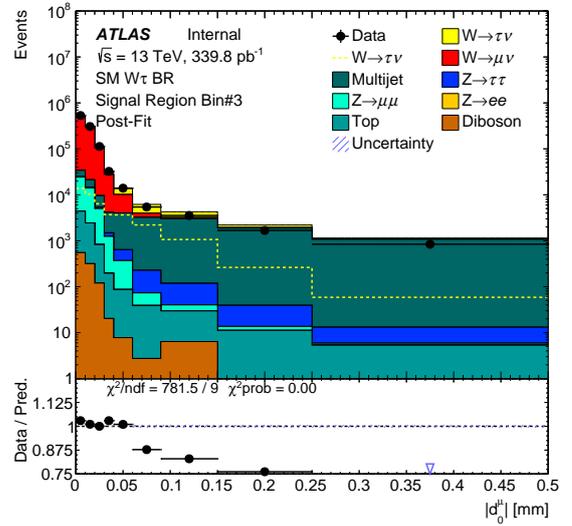
(a)



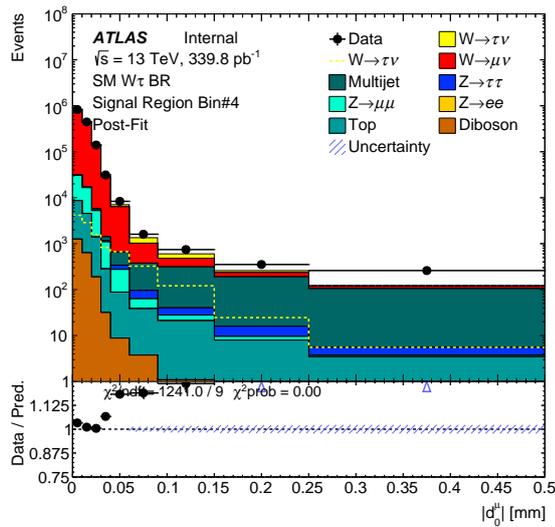
(б)



(в)

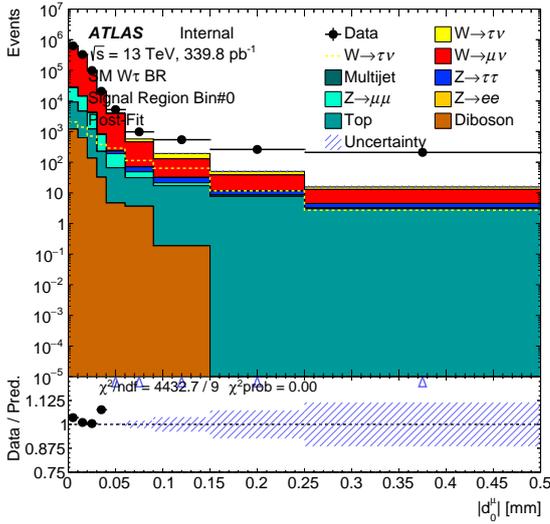


(г)

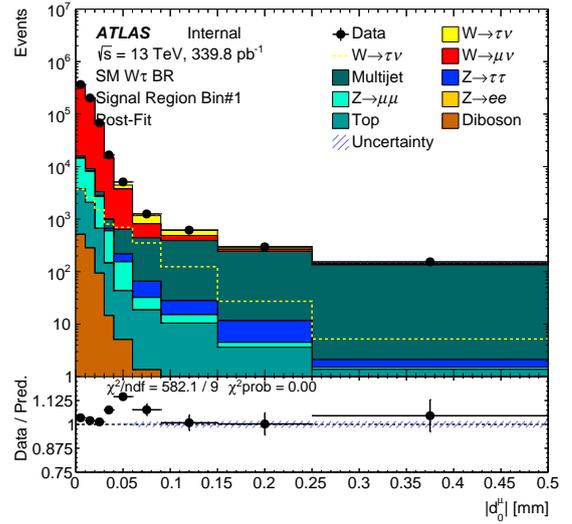


(д)

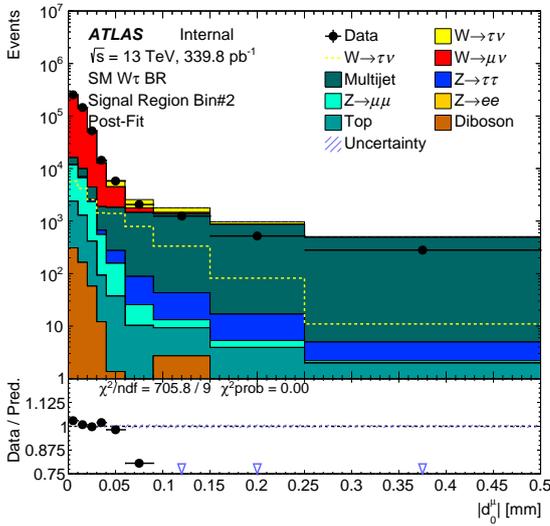
Рисунок 5.3.2 – Разбиение переменной  $d_0$  по бинам переменной поперечной массы  $m_T$ . Первый бин  $40 < m_T < 41$  (а), второй бин  $41 < m_T < 44$  (б), третий бин  $44 < m_T < 51$  (в), четвертый бин переменной  $51 < m_T < 68$  (г), пятый бин переменной  $68 < m_T < 100$  (д).



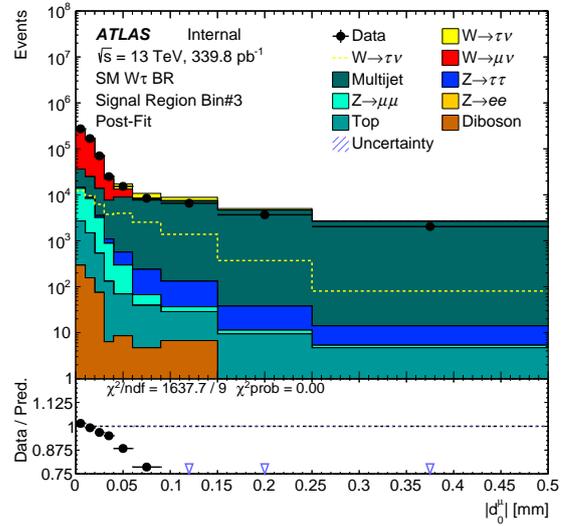
(a)



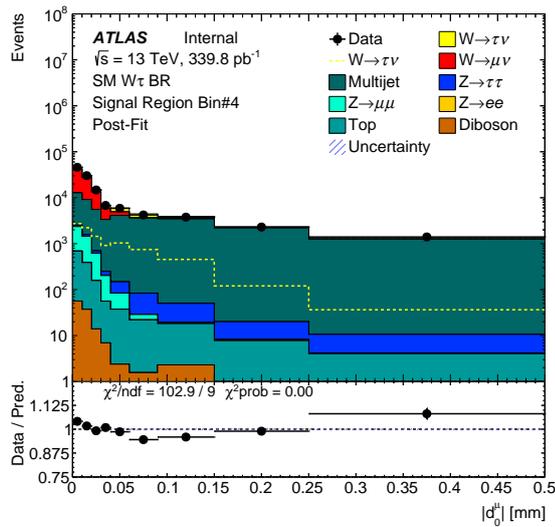
(б)



(в)



(г)



(д)

Рисунок 5.3.3 – Разбиение переменной  $d_0$  по бинам переменной BDT. Первый бин  $0.0 < BDT < 0.3$  (а), второй бин  $0.3 < BDT < 0.5$  (б), третий бин  $0.5 < BDT < 0.65$  (в), четвертый бин  $0.65 < BDT < 0.83$  (в), пятый бин  $0.83 < BDT < 1.0$  (г).

Был определен вклад (см. таблицу 5.3.1), который вносит группа систематических погрешностей в общую систематическую погрешность измерений. При использовании в фитировании поперечной массы наибольший вклад в систематическую погрешность вносят: погрешность оценки КХД фона, статистическая погрешность Монте-Карло. При использовании в фитировании новой переменной отклика классификатора BDT наибольший вклад в систематическую погрешность вносят те же группы систематических погрешностей. Малость вклада остальных групп погрешностей объясняется сокращением систематических погрешностей благодаря используемой стратегии анализа (измерении отношения отношений сечений).

Группа погрешностей	Контрольная переменная	
	$m_T$	BDT
	Погрешность	
Полная систематическая погрешность	0.022	0.027
Статистическая погрешность МК	0.014	0.017
Погрешность на оценку КХД	0.018	0.023
NormFactors	$< 10^{-3}$	0.003
cross_section	0.005	$< 10^{-3}$
lumi	0.008	$< 10^{-3}$
muIso_STAT	0.008	$< 10^{-3}$
muIso_SYS	0.008	0.011
muReco_STAT	0.007	0.001
muReco_SYS	0.008	$< 10^{-3}$
muTTVA_STAT	0.008	$< 10^{-3}$
muTTVA_SYS	0.008	$< 10^{-3}$
muTrig_STAT	0.008	0.009
muTrig_SYS	0.009	0.005

Таблица 5.3.1 – Вклад групп систематических погрешностей, входящих в полную систематическую погрешность.

Стоит отметить, что при разделении всех систематических погрешностей на группы, квадратичная сумма вкладов погрешностей от всех групп отличается от полной систематической погрешности. Это связано с корреляцией групп систематических погрешностей между собой.

# Заключение

В данной работе предпринята попытка повысить точность измерения лептонной универсальности с использованием методов машинного обучения для классификации прямых  $W \rightarrow \mu\nu$  и не прямых  $W \rightarrow \tau\nu \rightarrow \mu\nu\nu$  лептонных распадов  $W$  бозона. Работа выполнялась с использованием экспериментальных данных, набранных на эксперименте ATLAS во время второго сеанса набора данных, а также с использованием данных Монте-Карло.

Выполнен подбор подходящей архитектуры модели классификатора с использованием данных с большим  $\langle \mu \rangle$ . Исследовались модели классификаторов BDT и kNN. Критерием выбора оптимального классификатора выступал интеграл под ROC-кривой. Сделан выбор в пользу BDT классификатора использующего для классификации 70% обучающей выборки и архитектуру из 200 деревьев с максимальной глубиной в 5 ветвей. Значение интеграла под ROC-кривой для данной архитектуры составило 0.749.

Выбранная архитектура модели классификатора использовалась для тренировки модели и классификации событий для данных с низким значением  $\langle \mu \rangle$ . Значение интеграла под ROC-кривой для этих данных увеличилось до 0.795. Проведено сравнение отклика классификатора для экспериментальных данных и данных полученных методом Монте-Карло. С помощью дополнительного псевдо- $W$  региона показано, что наблюдаемые отклонения в отношениях экспериментальных данных к Монте-Карло симуляции в области от 0.6 до 1 вызваны неучтенным КХД фоном. Сделан вывод о оставшемся расхождении в области от 0.9 до 1. Причиной отклонения может являться как плохое моделирование Монте-Карло, так и неполная оценка КХД фона.

С использованием полученного отклика классификатора BDT на данные с низким значением  $\langle \mu \rangle$ , выполнена оценка точности статистического анализа лептонной универсальности. Проведена оптимизация бинов

переменной отклика модели BDT. Оптимальным бинингом для переменной отклика классификатора BDT является: [0.0, 0.3, 0.5, 0.65, 0.83, 1.0]. Оценка точности измерения отношения сечений проводилась с использованием данных Азимова. Учитывались экспериментальные систематические погрешности, в том числе систематические погрешности на оценку КХД фона. Результат составил:

$$\mu_{sig} = 1.000^{+0.015}_{-0.015}(stat)^{+0.027}_{-0.027}(syst)$$

Для сравнения выполнен аналогичный статистический анализ, но уже с использованием переменной поперечной массы  $m_T$ , реконструируемого  $W$  бозона. Полученный результат не отличается от результата приведенного выше, за исключением незначительного улучшения систематической погрешности измерения. Это означает, что используемый классификатор не получает новой информации для классификации событий из кинематических переменных в сравнении с информацией о кинематике событий, заложенной в переменной поперечной массы.

В процессе выполнения данной работы было выполнено ознакомление с пакетом ROOT, библиотекой машинного обучения TMVA. Получены начальные сведения о Стандартной модели, лептонной универсальности, экспериментах по проверке лептонной универсальности, эксперименте ATLAS и машинном обучении. Также были изучены следующие пакеты для проведения анализа данных: xTauReader, HAPPY, TRExFitter. Выполнен ряд работ по модернизации xTauReader фреймворка. За счет данных улучшения удалось значительно повысить вычислительную производительность на этапе классификации данных.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Electroweak Measurements in Electron-Positron Collisions at W-Boson-Pair Energies at LEP / S. Schael [et al.] // Phys. Rept. — 2013. — Vol. 532. — P. 119–244. — arXiv: [1302.3415 \[hep-ex\]](#).
2. SU(5) Unification with TeV-scale Leptoquarks / P. Cox [et al.] // JHEP. — 2017. — Vol. 03. — P. 035. — arXiv: [1612.03923 \[hep-ph\]](#).
3. Quark flavor transitions in  $L_\mu - L_\tau$  models / W. Altmannshofer [et al.] // Phys. Rev. D. — 2014. — Vol. 89. — P. 095033. — arXiv: [1403.1269 \[hep-ph\]](#).
4. *Емельянов В. М.* Стандартная модель и её расширения. — НМ. : Физматлит, 2007. — с. 584.
5. Precision measurement of the mass of the  $\tau$  lepton / M. Ablikim [et al.] // Phys. Rev. — 2014. — Vol. D90, no. 1. — P. 012001. — arXiv: [1405.1076 \[hep-ex\]](#).
6. Test of lepton universality in beauty-quark decays / R. Aaij [et al.]. — 2021. — arXiv: [2103.11769 \[hep-ex\]](#).
7. Measurement of Branching Fractions and Rate Asymmetries in the Rare Decays  $B \rightarrow K^{(*)}l^+l^-$  / J. P. Lees [et al.] // Phys. Rev. D. — 2012. — Vol. 86. — P. 032012. — arXiv: [1204.3933 \[hep-ex\]](#).
8. Test of lepton flavor universality and search for lepton flavor violation in  $B \rightarrow K\ell\ell$  decays / S. Choudhury [et al.] // JHEP. — 2021. — Vol. 03. — P. 105. — arXiv: [1908.01848 \[hep-ex\]](#).

9. Search for lepton-universality violation in  $B^+ \rightarrow K^+\ell^+\ell^-$  decays / R. Aaij [et al.] // Phys. Rev. Lett. — 2019. — Vol. 122, no. 19. — P. 191801. — arXiv: [1903.09252 \[hep-ex\]](#).
10. LEP design report. — Geneva : CERN, 1984. — Copies shelved as reports in LEP, PS and SPS libraries.
11. A Combination of preliminary electroweak measurements and constraints on the standard model. — 2005. — arXiv: [hep-ex/0511027](#).
12. Test of the universality of  $\tau$  and  $\mu$  lepton couplings in  $W$ -boson decays from  $t\bar{t}$  events with the ATLAS detector / G. Aad [et al.]. — 2020. — arXiv: [2007.14040 \[hep-ex\]](#).
13. The ATLAS Experiment at the CERN Large Hadron Collider / G. Aad [et al.] // JINST. — 2008. — Vol. 3. — S08003.
14. A precision measurement of the  $W$  boson decay branching fractions in  $pp$  collisions at  $\sqrt{s} = 13$  TeV : tech. rep. / CERN. — Geneva, 2021.
15. The CMS Experiment at the CERN LHC / S. Chatrchyan [et al.] // JINST. — 2008. — Vol. 3. — S08004.
16. TMVA - Toolkit for Multivariate Data Analysis / A. Hoecker [et al.]. — 2007. — arXiv: [physics/0703039 \[physics.data-an\]](#).
17. ATLAS xTauReader software documentation. — URL: [https://gitlab.cern.ch/Wlep\\_BR/xTauReader/blob/master/doc/README.md](https://gitlab.cern.ch/Wlep_BR/xTauReader/blob/master/doc/README.md).
18. ATLAS TReXFitter software documentation. — URL: [https://gitlab.cern.ch/atlas-wbr-lowmu/TReXFitter\\_SMWBR/-/blob/master/README.md](https://gitlab.cern.ch/atlas-wbr-lowmu/TReXFitter_SMWBR/-/blob/master/README.md).
19. Muon reconstruction and identification efficiency in ATLAS using the full Run 2  $pp$  collision data set at  $\sqrt{s} = 13$  TeV / G. Aad [et al.]. — 2020. — arXiv: [2012.00578 \[hep-ex\]](#).
20. ATLAS HistMaker software documentation. — URL: <https://gitlab.cern.ch/atlas-wbr-lowmu/HistMaker/-/blob/master/README.md>.
21. *Sjostrand T., Mrenna S., Skands P. Z.* A Brief Introduction to PYTHIA 8.1 // Comput. Phys. Commun. — 2008. — Vol. 178. — P. 852–867. — arXiv: [0710.3820 \[hep-ph\]](#).

22. Event Generation with Sherpa 2.2 / E. Bothmann [et al.] // SciPost Phys. — 2019. — Vol. 7, no. 3. — P. 034. — arXiv: [1905.09127](https://arxiv.org/abs/1905.09127) [hep-ph].
23. Combined Performance (CP) Groups. — URL: <https://twiki.cern.ch/twiki/bin/view/AtlasProtected/AtlasPhysics>.
24. *Bothmann E., Schönherr M., Schumann S.* Reweighting QCD matrix-element and parton-shower calculations // Eur. Phys. J. C. — 2016. — Vol. 76, no. 11. — P. 590. — arXiv: [1606.08753](https://arxiv.org/abs/1606.08753) [hep-ph].
25. Multi-jet background in low-pile-up runs taken in 2017 and 2018 / T. Xu [et al.]. — 2019.
26. ATLAS WBR analysis with low mu. — URL: <https://twiki.cern.ch/twiki/bin/viewauth/AtlasProtected/Vtaus13TeV>.
27. Measurement of the ratio of branching fractions of W-boson decays to  $\tau$  and light leptons in  $t\bar{t}$  events with the ATLAS detector : tech. rep. / J. A. McFayden [et al.] ; CERN. — Geneva, 2019.